Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

ICTIMES - INTERNATIONAL CONFERENCE ON EMERGING TECHNOLOGIES IN COMPUTER SCIENCE (ICETCS)

S.NO	Title of the Paper	Page No.
IC18CS1	A Survey on Reversible Data Hiding in Digital Image -- *Karpagaselvi Subramanian Vairachilai S, Mearg Enqubahri*	1
IC18CS2	Chrome Extension Webixnary: A Dictionary and a Page Review Tool -- *Srishti Singh, Vairachilai S, Gowtham*	7
IC18CS3	A Way to Secure the Data in Cloud Data Storage by Using Cloud Data Compression Mechanism -- *Shaik Saidhbi, Komati Thirupathi Rao*	15
IC18CS4	Survey on Face Recognition Techniques and Approaches -- *Nikita Mor, Vairachilai S, Brahma Naidu Kakarla*	23
IC18CS5	Improved Disease Diagnosis in Healthcare Using FDCT and RDWT Based Medical Fusion Technique -- *Lakshmi Devasena C*	27
IC18CS6	A New Text Mining Approach for Web Content Search -- *Thayyaba Khatoon Mohammed,Chandrasekar V , A Govardhan*	32
IC18CS7	A Comprehensive Review on Density-Based Clustering Algorithm in Data Mining -- *R Sujatha , Dr.V.Chandrasekhar*	36
IC18CS8	Design of Chaotic Behavior for Programmable Cellular Automata Based Symmetric Key Encryption Algorithm -- *Dr Chandrasekar V, Dr Shanthi S*	42
IC18CS9	IoT Based Smart Village for the Rural Development -- *Dr.J.Gladson Maria Britto , Dr.V.Bhoopathy*	46
IC18CS10	Evaluating Reliable Node in Wireless Sensor Network Using ERLEN Model *M.S. Nidhya, Dr. Kannagi A Dr.Chinnaiyan R*	52
IC18CS11	Tackling Data Science for Aliment forecasting in Health Care -- *Nagappa Gandla, Praveen Kumur G, Ravi R*	57
IC18CS12	Survey on Women Safety Based on IoT -- *Thamaraiselvi K, Rinesh S, Kalpana A*	66
IC18CS13	A New Framework for Video Data Retrieval Using Hierarchical Clustering Technique	71

ISBN 978-81-939386-1-4

	-- *D.Saravanan*	
IC18CS14	Video Data Retrieval Using Image Color Histogram Technique -- *D.Saravanan*	76
IC18CS15	Solving Data Mining Challenge Through Link Mining -- *Jangam Ravali, Thati Sindhura, Madan Mohan*	81
IC18CS16	Secured Data - Mobile Cloud Computing -- *Suman Nayaka, Lagshetty Ananth Lakshmi, Arun K*	87
IC18CS17	Stock Hunt: Predictive Analysis Using Machine Learning -- *Shubhangi Urkude, Niteen Balpande*	96
IC18CS18	A TSM Mechanism to Enhance Node Trustworthiness and Using Asymmetric Key with Ant Intelligence to Secure DSR in Wireless Ad-Hoc Network -- *Veena Rani and Jakka Sushma*	101
IC18CS19	Spotting of Question Twin which has Semantically Similar -- *Vibha Lakshmi S, Vairachilai S, Brahma Naidu Kakarla*	108
IC18CS20	Fake Currency Detection Based on Fluorescence in HSV Color Space -- *Stalin Alex, P Subramanian, R Sravani*	112
IC18CS21	Effective Approaches for Securing Sensitive Data and Supporting Multi Data Applications in Cloud forensics Environment -- *Sonya A, Kavitha G*	115
IC18CS22	Cipher Text Policy for Attribute-Based Encryption in Cloud Data -- *Raja M, Dhanasekaran S*	119
IC18CS23	A Novel Path Inference Approach to Reconstructing the Packet Routing Paths in Dynamic and Large Scale Networks -- *Byreddy Madhavi, P. John Paul*	123
IC18CS24	Aggressive and Community Auditing System with Clear Agreement for Cloud Data -- *Vemula Naveena, P Subhashini*	128
IC18CS25	IOT In Rural India -- *Anil Kumar Vitkuri, L Shruthi, Ch Vijaya Kumari*	133
IC18CS26	Booster in High Dimensional Data Classification -- *Vishweshwar K, Dr. S Moses Dian*	138
IC18CS27	An Overlay Architecture for Throughput Optimal Multipath Routing -- *Mounika Sapa, Dr.Britto Gladson*	149
IC18CS28	SWEET: Serving the Web by Exploiting Email Tunnels -- *Devi Sravani, Dr.P. John Paul*	153
IC18CS29	Student Intervention System Using Machine Learning Techniques -- *Shubhangi Urkude, Kashitij Gupta*	164
IC18CS30	Robust Optimized DWT-SVD Based Watermarking -- *Swaraja Kuraparthi, Meenakshi K, Padmavathi Kora, Karuna G*	169

ISBN 978-81-939386-1-4

IC18CS31	Custom Queuing Method for Enhancing Security of a Packet Scheduling Process in Real Time Networks -- *Karuna G, Swaraja Kuraparthi, Meenakshi K, Padmavathi Kora*	176
IC18CS32	Nearest Keyword Set Search Queries on Multi-Dimensional Datasets -- *Aju P Benny, Dr.V. Bhoopathy*	181
IC18CS33	Achieving Efficient and Secure Data Acquisition for Cloud-Supported Internet of things in Smart Grid -- *Gudipudi Soundarya, Dr.V. Bhoopathy*	185
IC18CS34	Privacy and Integrity Preserving Top-K Query Processing for Two-Tiered Sensor Networks -- *Karrolla Bharath, Dr. M. Narayanan*	188
IC18CS35	Privacy Protection Based Access Control Scheme in Cloud-Based Services -- *P. Supriya, Dr. T. Sunil*	200
IC18CS36	Sentiment Embeddings with Applications to Applicant Sentiment Analysis -- *G. Priskilla, Dr.T. Sunil*	204
IC18CS37	Multi Disease Prediction Using Data Mining Techniques -- *Kanneboina Ashok, M. Mounika, Syed Rasheeduddin*	207
IC18CS38	Energy Efficient Resource Allocation in Cloud Computing -- *Dr.Sunil Tekale, Mattakoyya Aharonu, Ch.Vijaya Kumari*	212
IC18CS39	Web Innovation with IoT in Social Environment for Sharing Efficient Information -- *Kranthi M, Maheswari K Y, Arun K*	218
IC18CS40	A Survey on AI Driven Cloud Computing-Future Trends -- *Anitha Bejugama, Kota Divya Bharathi and Jagirdar Srinivas*	231
IC18CS41	Reorganization of DFA Based Prototype Matchers -- *Vengaiah Ch, Praveen Reddy K Aharon Mattakoyya*	236
IC18CS42	Protected and Flexible Multi-Keyword Search Model Over Encoded Cloud Data -- *Syed Mohammed Shafi*	240
IC18CS43	Oruta: Privacy-Maintaining Public Scrutinize for Shared Data in the Cloud -- *Tanna Mallika, Vaddi Devipriya Rudra Koteswaramma*	250
IC18CS44	Ontology Frame Work Based on Generalized Rough Set theory -- *Ala Prathap, Sudhapavani Kota, Kanaparthi Venkateswarlu*	265
IC18CS45	Providing Security for Outsourced Data in Cloud Storage -- *Riyaz Md, Veerender A , Malleswar Rao Ch*	271
IC18CS46	Different types of Machine Learning Techniques in Python Programming -- *Naveen Kumar R, Adepu Aravinda, Subhashini Peneti*	280
IC18CS47	Timing Enabled Proxy Re-Encryption Technique for E-Health Care Clouds in Keyword Search by Delegate -- *R Sai Hanuma Chandrika, J.Gladson Britto*	286
IC18CS48	Survey on Human Wearable Computing -- *Dr.Anandaraj S.P, Dr.Kirubakaran N*	300
IC18CS49	Classify the Peer-To-Peer Networks Videos and Video Servers Using AODE Algorithm -- *Narayanan M*	311

ISBN 978-81-939386-1-4

IC18CS50	A Neighbor Coverage Protocol with Broadcasting Mechanism for Dissemination in Mobile IoT Networks -- *P Sandhya Priyanka, V Tirumala, E Lingappa*	316
IC18CS51	Content Based Confidentiality Detection Method for Data Leakage Prevention -- *Subhashini Peneti, Padmaja Rani B*	319
IC18CS52	Identity-Based Encryption with Cloud Revocation Authority and its Applications -- *P.Pravalika Reddy, Prof.Kannagi*	327
IC18CS53	The Vehicle Network Architecture in Smart City With Blockchain Technology -- *Rajasekar Rangasamy, R.Jahnavi*	333
IC18CS54	A Survey on Detection and Prediction of Lung Cancer Using Different Algorithms -- *N.Sudhir Reddy, Dr.V Khanna,*	340

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A Survey on Reversible Data Hiding in Digital Image

Karpagaselvi Subramanian
School of Electrical and Computer Engineering Ethiopian Institute of Technology-Mekelle Mekelle University Ethiopia
karpaga.selvi@mu.edu.et

S.Vairachilai
Department of Computer Science and Engineering Faculty of Science and Technology, The ICFAI Foundation for Higher Education Hyderabad, India
vairachilai2676@gmail.com

Mearg Enqubahri
School of Electrical and Computer Engineering Ethiopian Institute of Technology-Mekelle Mekelle University Ethiopia

Abstract— The growth of multimedia technologies and the attractiveness of internet are dramatically increasing. All kinds of multimedia information like audio, video and images can be obtained freely and copied even edited then transmitted as one wishes. At he same time, this makes the integrity, reliability, and security of data transmitted under threat. The information security has always een a major concern. As one of the ways to solve the security problem is data hiding technology that embeds the secret data mperceptibly into the cover media by slightly modifying some of the cover elements. This paper discusses the basic principles of eversible data hiding (RDH) as a method of covert communication in order to understand how information is embedded in the raphical representation and covers the in-depth discussion on image compression.

Key words: Reversible data hiding, JPEG, Compression, discrete cosine transform, Quantization.

I. INTRODUCTION

Traditional data hiding techniques usually introduce rreversible distortion in the cover media which cannot be ecovered after secret data extraction. However, the istortion may be not allowed in the medical, military and egal fields. To overcome this problem, reversible data iding has been proposed as a new solution where both ecret data and the original cover media can be extracted and ecovered without any distortion. Moreover, data hiding echniques can be applied to other aspects such as mbedding patient's information into medical images and mbedding geographic information into remote sensing atellite images [1].

A flurry of research works has been contributed by esearchers about lossless data hiding scheme for Joint hotographic Experts Group (JPEG) images within in the ast few years. Some of the data hiding techniques on JPEG mages uses standard quantization table to embed data into n image while others embed data by modifying the uantization table. The JPEG images processing algorithms ary according to the purpose of embedding. Some uses ossless algorithms while others use reversible watermarking mbedding techniques. In the lossy algorithms, the mbedded data will be recovered where as in the lossless echniques both the message and the data will be recovered 2]. In reversible data hiding in JPEG, the discrete cosine ransform (DCT) coefficients is based on four general data iding approaches [2].

The first one is lossless compression based which was proposed by Fridrich [3] for secret data embedding. In this approach, a space is preserved by compressing the redundant component of the image. Since message capacity is too small, this method takes less attention of researchers.

The second approach is quantization table modification approach which was proposed by Fridrich [4] to authenticate digital image. Later, Wang et al. [5] proposed a method which is based on modifying the quantization table and quantized DCT coefficients. In this approach, some elements of the quantization table are divided by an integer while the corresponding quantized DCT coefficients are multiplied by the same integer and added by an adjustment value to make space for embedding the data. This result shows an improvement on the Peak Signal to Noise Ratio (PSNR) value over previous works. This approach is basically works by prepossessing the quantized DCT coefficients to create space for secret data embedding. This is done by making the selected DCT coefficients a multiple of an integer k. If D is quantized DCT coefficients then D' (preprocessed DCT coefficients) is multiple of D and k, where k c {2, 3 …}. The corresponding quantization table entry again divided by k to compensate multiplication side effect on image distortion. The remainders of the quantization table division padded to the main message as side information so that decoder will use it to construct the original values. The binary secret message is changed into k's based number format is another part of the method to have high capacity embed ability efficiency. Finally, according to the embedding algorithm encoder can embed secret data in the modified DCT values. Using the pre-defined integer, decoder can easy apply modulo operation on modified DCT coefficients to get the embedded secret

message, divide DCT coefficients and multiply quantization table entries to get the original DCT and quantization table values respectively.

The third method is based on modification of Huffman table which was proposed by Mobasseri et al. [6]. In this method, data embedding is performed by mapping variable length code (VLC) to an unused VLC. The Qian and Zhang's scheme [2] is another research work related to data embedding on JPEG images. The work can be seen from the point of JPEG bit stream structure and the scheme perspectives. From the JPEG bit stream structure perspective, an input gray-scale image is compressed into the JPEG format by the JPEG encoder and saved in the form of bit stream. It starts with an Start of Image (SOI) marker and ends with an End of Image (EOI) marker. The Huffman data is used to generate the Huffman table. The scheme is based on Mobasseri et al.'s method which works by flipping one or more bits of the non-zero AC Coefficients. This work also improved of mobasseri et al.'s method [6], but still the payload is too small.

The fourth method is based on modifying the quantized DCT coefficients. This coefficients modification approach is proposed by Chang et al. [7] and Xuan et al. [8]. Sakai et al. [9] improved Xuan et al. [8] method by block selection approach and get better visual quality of the image. Li et al. [10] proposed reversible data hiding scheme on JPEG images based on smaller DCT value selection and slightly modification of quantization table. Based on histogram shift (HS) strategy, Lin et al. [11] discussed high capacity reversible data hiding for JPEG image. They used different block sizes to get high embedding capacity. Celik et al. [12] proposed lossless data hiding in JPEG DCT coefficients least significant bit (LSB). In [12], the DCT value which has high distortion is padded as side information to the main secret message. Wang et al. [1], Xuan et al. [8] and Sakai et al. [9] are done the best research works on reversible data hiding for JPEG images based on quantization table and DCT values modification scenario. Although the experimental result of [1] show better PSNR, the file size increate a lot because of multiplication operation. If D = 15 and k = 3, then D' becomes 45. The memory space used to store D' is 2 bit more compared to D. Shifts the quantized DCT coefficient histogram with an optimum searching strategy. This optimum strategy used to achieve good performance. In order to make data embedding unperceivable and the visual quality of marked image higher when embedding a certain amount of data, only lower and mid-frequency DCT coefficients are chosen to embed data in the embedding process. Sakai et al. [9] improved Xuan et al.'s scheme in 2008.This method yielded better image quality based on new block selection strategy.

II. Data Hiding

Data hiding is a group of techniques used to keep a secure data in a host media with small deterioration in host that can be used to extract the same data in later on process. Steganography is one such pro-security innovation in which secret data is embedded in a cover [13]. This method fall into reversible data hiding. In reversible data-hiding, the information bits are inserted by modifying the host signal that will enable the lossless restoration of the original host signal after extracting the embedded information. However, the possibility of recovering the exact original image is a desirable property in many fields such as legal, medical, and military imaging.

The documents like bank checks are scanned & protected with an authentication scheme based on a reversible data hiding and sent through the Internet. In most cases, the watermarked documents will be sufficient to distinguish unambiguously the contents of the documents. However, if any uncertainty arises, the possibility of recovering the original unmarked document is very interesting [14]. The block diagram of a basic reversible data hiding system shown in Figure 1. The existing reversible data hiding schemes are mostly fragile in nature. The two important properties of reversible data hiding are imperceptibility and embedding capacity. The imperceptibility is the measure of similarity between stego and the cover image. The embedding capacity is the measure of the maximum number of information bits that can be embedded in the cover image. The performance of a reversible data hiding technique is evaluated on the basis of imperceptibility and embedding capacity.

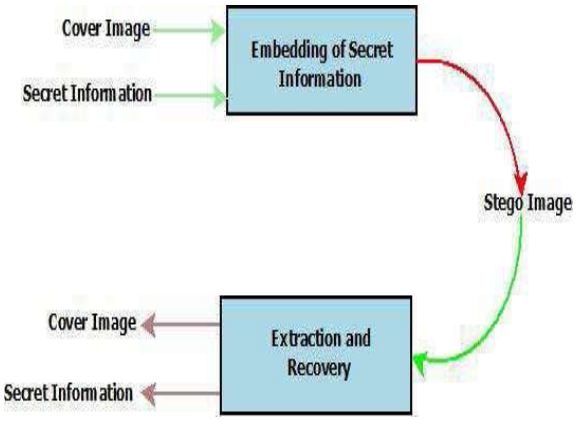


Figure 1: Block diagram of basic reversible data hiding system [[13]

III. Reversible Data Hiding Techniques

Data hiding techniques are characterized by the embedding and extraction techniques. The reversible data hiding techniques classified into compression based, histogram modification based, quantization based, and expansion based techniques. The RDH techniques are getting more popular because of its increasing sensitive applications areas such as military communication, healthcare, and law-enforcement [13].

In the medical imagery field, if a medical image is modified using conventional watermarking, the small change may affect the interpretation significantly and a physician may make a wrong diagnosis. In military application, the changes due to embedding of secret information can

substantially alter the cover image and therefore, the decision taken may cost considerably. Consequently, there is a strong need to restore the cover work to its original form. The reversible data hiding known as lossless data hiding that allows full extraction of the embedded information along with the complete restoration of the cover work. The reversible data hiding can be considered as a special case of steganography.

A. Compression based RDH

In order to recover the original image, the information is stored along with the secret data. In case of RDH, additional data needs to be embedded. In addition, it needs more space compared to conventional data hiding for data embedding. A simple approach will compress a part of cover image for embedding data. Several reversible data hiding schemes are reported using this approach [15]. In this approach, image is first transformed into frequency domains by taking integer wavelet transform and then dividing into blocks. The compression operation is performed on each coefficient of a block whose value is greater than a certain threshold. Each block has its own threshold value.

B. Histogram Modification Based RDH

In this type of RDH, histogram of the image in pixel/transform domain is calculated. Then, based on the histogram plot the schemes try to embed data by changing the plot of original histogram. This is based on the concept of histogram shifting. Ni et al. [16] proposed a reversible data hiding method which shifts slightly the part of the histogram between the maximum point (peak point) and the minimum point to the right side by one pixel value to create an empty bin besides the maximum point for hiding an input message. This method yields hiding capacities and provides higher qualities in stego-images. The knowledge of the maximum point and the minimum point of the histogram is necessary for retrieving the hidden data and restoring the stego-image losslessly to the original state. In this method, P is the value of peak point and Z is the value of zero point. The range of the histogram P + 1 & Z – 1 is shifted to the right-hand side by 1. Once a pixel with value P is encountered, if the message bit is set as 1, the pixel value is increased by 1 otherwise no modification is needed.

The data extraction is the reverse of the data hiding process. In this approach, first, image is divided into several blocks of neighboring pixels and then each block is split into two zones. Finally, the corresponding histograms are calculated. A bin is shifted in accordance with the value of corresponding secret bit. If the bit is 1, shift the lowest bin to the highest one and downgrade other bins. If the bit is 0, upgrade each bin and shift the highest bin to the lowest bin. Ni et al. [16] utilizes zero or minimum point of histogram. If the peak is lower than the zero or minimum point in the histogram, it increases pixel values by one from higher than the peak to lower than the zero or minimum point in the histogram. While embedding, the whole image is searched. Once a peak-pixel value is encountered, if the bit to be embedded is '1' the pixel is added by 1, else it is kept intact. Alternatively, if the peak is higher than the zero or minimum point in the histogram, the algorithm decreases pixel values by one from lower than the peak to higher than the zero or minimum point in the histogram and to embed bit '1' the encountered peak-pixel value is subtracted by 1.

C. Difference Expansion Based RDH

Tian [17] proposed a high quality reversible watermarking method with high capacity based on difference expansion. The pixel differences are used to embed the data because of high redundancies among the neighboring pixel values in natural images. The neighboring pixel values are calculated during the embedding process and the changeable bits differences are determined. Some differences are chosen to be expandable by 1-bit, the changeable bits increases, concatenated bit-stream of compressed original changeable bits, the location of expanded difference numbers (location map) and the hash of original image (payload) is embedded into the changeable bits of difference numbers in a pseudo random order. This used the inverse transform to have the watermarked pixels from resultant differences. This method is good in video and audio but it suffers significant degradation of visual quality due to bit-replacements of gray scale pixels. This method depends on the smoothness of natural image. It cannot be applied to textured image [18].

D. Quantization based RDH

Quantization based data hiding techniques are robust. However, the reversible quantization based data hiding approaches are mostly fragile in nature. In this method, embedding a message into a host signal. This refers to quantization of the host signal by a quantizer picked from an ensemble of quantizer where each quantizer is associated with a message letter/index. One of the relevant approaches is Sequential Quantization Strategy (SQS) for data embedding. The SQS makes the modulation of a pixel value dependent on the previous pixels. A reversible data hiding method is used with SQS to make it more suitable for the authentication purposes. This algorithm ensures the recovery of cover image [14] for medical imaging systems.

E. DCT Based Data Hiding

For the past few years, a standardization effort known by the acronym JPEG [19, 6, 20] has been working toward establishing the first international digital image compression standard for continuous-tone (multilevel) still images both grayscale and color. The discrete cosine transform is a general orthogonal transform for digital image processing and signal processing. The advantages of this method are small bit error rate, good information integration ability, good synthetic effect of calculation complexity and high compression ratio [21]. This scheme also requires a preprocessing phase because the secret data are hidden in the DCT coefficients of a cover image. In preprocessing phase, first the JPEG image partitioning into non-overlapping blocks of 8 x 8 blocks. Second, the DCT of each block are computed and the resultant DCT coefficient matrix is quantized using a standard quantization table. Finally, the inverse DCT of quantized coefficient matrix are evaluated and the final JPEG image is obtained after rounding the

values [4]. The key steps of JPEG compression process [19, 22, 20] is shown in Figure 2. The color image can be approximately regarded as multiple grayscale images. The JPEG encoder (DCT based encoder) consists of three parts: (i) DCT (ii) Quantizer (iii) Entropy coder. First, the original image is divided to 8x8 non-overlapping blocks. Second, each block will be transformed from spatial domain to frequency domain using two-dimensional DCT function. In the quantization phase, the obtained DCT coefficients are divided by the predetermined quantization table. The quantized DCT coefficients are arranged in a zigzag order. After scanning the DCT coefficients pre-compressed using the differential pulse code modulation (DPCM) on DC coefficients and run length encoding (RLE) on AC coefficients. Finally, the symbol string is Huffman coded to obtain the final compressed bit-stream. After pre-pending the header, the final JPEG file is obtained.

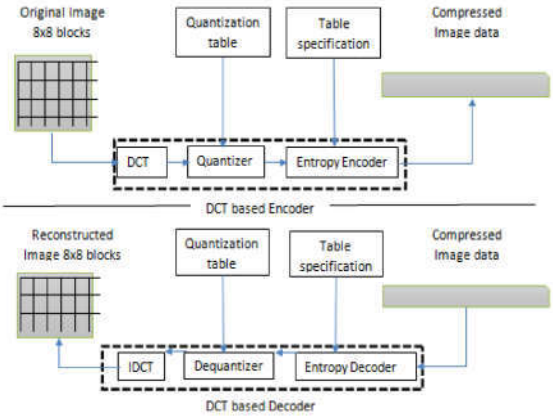


Figure 2: The block diagram of JPEG compression process

IV. JPEG COMPRESSION

JPEG [23] format is one of the most popular formats and widely used on various platforms [1]. For instance, an uncompressed 640 * 480 pixel 24-bit color image would require 900 kilobyte (KB), whereas a JPEG version of the same image can be compressed to 150 KB[24]. In data hiding there are two common methods of embedding: (i) Spatial domain method [23] in which messages are inserted into the LSBs of image pixels (ii) Variable bit LSB. In LSB substitution, images directly embedded into the pixels of the cover image by using 1bit-LSB, 2bit-LSB, Variable bit LSB. In the frequency-domain [16, 23] the secret image is first transformed to frequency-domain, and then the messages are embedded in the transformed coefficients. Transform embedding methods are more robust than the spatial embedding methods which are susceptible to image-processing type of attacks. The discrete cosine transform [16] is a widely used tool for frequency transformation. If we apply JPEG images to data hiding, the stego-image will not easily draw attention of suspect. The DCT methods gives minimum distortion between the original cover and stego images while Walsh gives the maximum distortion. DCT is the most efficient transform for embedding [23].

A. Co-efficient Quantization

The DCT output matrix takes more space to store than the original matrix of pixels. The input to the DCT functions consists of eight bit pixels values. The values range from a low -1024 to a high 1023, occupying eleven bits. To discard an appropriate amount of information, the encoder divides each DCT output value by a quantization value and rounds the result to an integer. A quantization value of sixteen produces a quantized DCT co-efficient with seven bit precision. If the quantization value is large then the more data is lost. The actual DCT value is represented less accurately. Quantization is simply the process of reducing the number of bits needed to store an output value by reducing the number of bits to represent integer results. Each of the 64 positions in the 8x8 matrix has its own quantization value with the higher order terms being quantized more heavily than the low-order terms. Thus, the higher-order terms have larger quantization values. In the quantization table/ matrix, for every element position in the DCT matrix has a corresponding quantum value. The quantum value indicates what the step size is going to be for that element in the compressed rendition of the picture with values ranging from 1 to 255. The elements to the picture will be encoded with a small step size, the size one offering the most precision. The values can become higher as they move away from the origin. The formula for quantized is as follows:

$$\text{Quantized Value}(u,v) = \frac{\text{DCT}(u,v)}{\text{Quantum}(u,v)}$$

JPEG image compression is based on the DCT, which is one of the basic building blocks for JPEG compression [19, 22]. The important aspect of DCT for JPEG compression is the ability to quantize the DCT coefficients using visually weighted quantization values. JPEG standard does not specify the DCT quantization table that everybody (researchers) can use their own quantization table by scaling the table elements using different quality factor [19, 22]. The different quantization table used the different image quality and compression ratio. The most known quantization table used by many researchers is shown in Table 1.

A scaling factor (SF) is computed with Equation (2) with the input Quality factor (QF) which is a positive integer number up to 100, QF = {1, 2… 100} is used to scale the quantization table values [1]. The scaled quantization table is obtained using Equation (1). The scaled quantization matrix[1] is then rounded and clipped to have positive integer values ranging from 1 to 255 using Equation (3).

The 8 x 8 block of DCT coefficients is compressed by quantization. The quantization is achieved by dividing each element in the DCT coefficient block by the corresponding value in the quantization matrix, and the result is rounded to the nearest integer. A useful feature in this process is the image compression. The quality is obtainable through selection of specific quantization table. The DCT block consists of 64 DCT coefficients. The top-left coefficients (0,

0) the original image block, which is called Direct current (DC) coefficient. As we move away from the F (0, 0) in all directions the DCT coefficients are Alternative current (AC) coefficients.

$$SF = \begin{cases} \frac{5000}{QF}, & if\ QF < 50 \\ 200 - (2 * QF), & if\ Qf \geq 50 \end{cases} \quad (2)$$

$$Q_{scaled}(i,j) = MAX\left(round\left(Q_{initial}(i,j) * \frac{SF}{100}) \right) \right| \quad (3)$$

TABLE 1. Standard Quantization Table (Q_{stan})

10	7	6	10	14	4	31	37
7	7	8	11	16	35	36	33
8	8	10	14	24	34	41	34
8	10	13	17	31	52	48	37
11	13	22	34	41	65	62	46
14	21	33	38	49	62	68	55
29	38	47	52	62	73	72	61
43	55	57	59	67	60	62	59

TABLE 2: Scaled Quantization Table (QF=70)

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

The 8 x 8 block of DCT coefficients is compressed by quantization. The quantization is achieved by dividing each element in the DCT coefficient block by the corresponding value in the quantization matrix, and the result is rounded to the nearest integer. A useful feature in this process is the image compression. The quality is obtainable through selection of specific quantization table. The DCT block consists of 64 DCT coefficients. The top-left coefficients (0, 0) the original image block, which is called Direct current (DC) coefficient. As we move away from the F (0, 0) in all directions the DCT coefficients are Alternative current (AC) coefficients.

A. Zigzag sequence

After the quantization, the next step in the JPEG algorithm is to entropy encoding of the quantized DCT coefficients. The DC coefficients are a measure of the average value of the 64 image samples. Because there is usually a strong correlation between the DC coefficients from the DC term of the previous block in the encoding order. The DC coefficients frequently contain a significant fraction of the total image energy. Finally, all quantized AC coefficients are ordered into a zigzag sequence that shown in Table 2. This ordering of coefficient helps to facilitate entropy encoding by placing the low frequency coefficients (which are more highly to be non-zero) before high frequency components.

V. CONCLUSION

Reversible data hiding techniques are used to hide secrete information inside cover image, as well as to recover the cover image after extraction of secrete message. RHD can be applied on encrypted image so that others do not know the original content. In this paper, we have reviewed several methods to understand how information is embedded in the graphical representation. A survey on various reversible data hiding techniques and image compression are discussed. In future, it is essential to develop an efficient method that provides data embedding and recovery without any distortion and provides better security

REFERENCES

[1] Wang, Kan, Zhe-Ming Lu, and Yong-Jian Hu, "A high capacity lossless data hiding scheme for JPEG images," Journal of Systems and Software, Vol. 86, No. 7, pp. 1965-1975, 2013.

[2] Qian, Zhenxing, and Xinpeng Zhang, "Lossless data hiding in JPEG bitstream," Journal of Systems and Software, Vol. 85, No. 2, pp. 309–313, 2012.

[3] Fridrich, Jessica, Miroslav Goljan, and Rui Du, "Lossless data embedding for all image formats," International Society for Optics and Photonics, Vol. 4675, pp. 572-583, 2002.

[4] Fridrich, Jiri, Miroslav Goljan, and Rui Du, "Invertible authentication watermark for JPEG images," Proc. Information Technology: Coding and Computing, IEEE Xplore, pp.223-227, DOI:10.1109/ITCC.2001.918795, 2001.

[5] Wang, Kan, Zhe-Ming Lu, and Yong-Jian Hu, "A high capacity lossless data hiding scheme for JPEG images," Journal of Systems and Software, Vol. 86, No. 7, pp. 1965-1975, 2013.

[6] Mobasseri, Bijan G, Robert J.Berger, Michael P.Marcinak, and Yatish J.NaikRaikar. "Data embedding in JPEG bitstream by code mapping," IEEE Transactions on Image Processing, Vol. 19, No. 4, pp. 958-966, 2010.

[7] Chin-Chen Chang, Chia-Chen Lin, Chun-Sen Tseng, and Wei-Liang Tai,"Reversible hiding in DCT-based compressed images," Information Sciences, Vol. 177, No. 13, pp. 2768–2786, 2007.

[8] Xuan, Guorong, Yun Q. Shi, Zhicheng Ni, Peiqi Chai, Xia Cui, and Xuefeng Tong, "Reversible data hiding for JPEG images based on

histogram pairs," Image Analysis and Recognition, Lecture Note in computer science, Vol. 4633, pp. 715-727, 2007.

[9] Sakai, Hirofumi, Minoru Kuribayashi, and MasakatuMorii, "Adaptive reversible data hiding for JPEG images," Information Theory and Its Applications, IEEE International Symposium, pp. 1-6, DOI: 10.1109/ISITA.2008.4895529, 2008.

[10] Li, Qiming, Yongdong Wu, and Feng Bao, "A reversible data hiding scheme for JPEG images," In Advances in Multimedia Information Processing-PCM, vol. 6297, Lecture Note in computer science, pp. 653-664. 2010.

[11] Lin, Yih-Kai, "High capacity reversible data hiding scheme based upon discrete cosine transformation." Journal of Systems and Software, Vol. 85, No. 10, pp. 2395–2404, 2012.

[12] Celik M.U, Gaurav Sharma, Murat Tekalp A, and Eli Saber, "Lossless generalized-LSB data embedding," IEEE Transaction on Image Processing, Vol. 14, No. 2, pp. 253-266, 2005.

[13] Pooja Prabhakar Petkar and P. S. Nalwade, "A Survey on Reversible Data Hiding Techniques," International of Computer Technology & Applications,Vol 5,pp. 897-903, 2014.

[14] Vaishnavi, V. and Kuechler, W, "Design Science Research in Information Systems" Auerbach Publications Boston, MA, USA, 2007.

[15] Ankita Meenpal and Shital S Mali, "Survey of Recent Reversible Data Hiding Schemes," International Journal of Advance Foundation and Research in Computer, Vol` 1, pp.2348 - 4853, 2014.

[16] Ni, Zhicheng, Shi Qing, Nirwan Ansari and Wei Fang Su, "Reversible data hiding," IEEE Transaction on Circuits and Systems for Video Technology, Vol. 16, No. 3, pp. 354-362, 2006

[17] Tian, Jun, "Reversible data embedding using a difference expansion," IEEE Transaction on Circuits System for Video Technology, Vol. 13, No. 8, pp. 890-896, 2003.

[18] Mohammad Awrangjeb, "An Overview of Reversible Data Hiding," Conference: International Conference on Computer and Information Technology, pp 75-79, 2003.

[19] Pennebaker, William B., and Joan L. Mitchell. JPEG: Still image data compression standard. Springer Science & Business Media, 1993.

[20] Suhad Shakir Zahereel I. and Abdul Khalib," Survey On Recent Digital Image Steganography Techniques" Journal of Theoretical and Applied Information Technology, Vol. 66 No.3,2014.

[21] Blossom Kaur, Amandeep Kaur, and Jasdeep Singh, "Steganographic approach for hiding image in dct domain," International journal of advances in engineering & technology, pp. 2231-1963, 2011.

[22] Wallace, Gregory K, "The JPEG still picture compression standard," Communications of the ACM, Vol. 34, No. 4, pp. 30-44, 1991.

[23] Xiaochao Qu, and HyoungJoong Kim, "Pixel-based pixel value ordering predictor for high-fidelity reversible data hiding," Signal Processing, Vol. 111, pp. 249-260, 2015.

[24] Jesse D. Kornblum, "Using JPEG quantization tables to identify imagery processed by software," Digital investigation Vol.5, pp. 21-25, 2008.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Chrome Extension – Webixnary: A Dictionary and a Page Review Tool

Srishti Singh
Department of Computer Science Engineering, Faculty of Science and Technology, The ICFAI Foundation for Higher Education
Hyderabad, India
srish3001@gmail.com

S. Vairachilai
Department of Computer Science Engineering, Faculty of Science and Technology, The ICFAI Foundation for Higher Education
Hyderabad, India
vairachilai2676@gmail.com

M. Gowtham
Department of Computer Science Engineering, Faculty of Science and Technology, The ICFAI Foundation for Higher Education
Hyderabad, India
chmgowtham@gmail.com

***Abstract* —The growing popularity and usage of the internet, every internet user is using one or more web browsers for surfing the net. But web browsers are limited to their basic functionality. So browser extensions come into picture which are small and simple applications that add functionalities to a given web browser. This paper reports a small-scale empirical study on the popularity and usage of different web browsers through a survey and how familiar participants in the survey are with browser extensions. For this study, 200 participants answered questions like how many web browsers do they use, which web browser they prefer over the others and how many browser extensions they use per browser, etc. This paper also reviewed 60 Chrome extensions which were similar to the application developed in the paper and discussed the similarities and differences among them also discussed. In this paper, we studied the significance of these browser extensions and the ease of developing one for a given web browser. As a proof of concept, we present Webixnary, a Chrome extension which helps users get the definition of a selected word on a webpage just through a popup and without having to leave the page itself. Webixnary also has a summary-page review functionality which is not found in any of the extensions studied in this research or to our knowledge in the area.**

Keywords— Web browsers, Chrome extension, Web technologies.

I. INTRODUCTION

As the number of internet users is growing every moment, the amount of time these users spent on the web is also increasing. Web browser being the portal to everything that is there on the Web, it's only logical to imply that each of these users uses one or more web browsers for browsing through the Web. But web browsers are limited to its basic functionality. So browser extensions are used to extend that limited functionality. Browser extensions or also known as add-ons are small software applications that add functionality to a web browser. They can modify, control and alter web pages. The following are few different reasons we use browser extension for [28]: (i) To integrate with other applications. For example, *Evernote,* an extension allows users to easily clip websites and save them to their *Evernote* account. (ii) To add additional features to the browser. For example, the *JoinTabs* extension for Chrome gives a button to click to combine all the Chrome tabs from multiple windows into a single window. (iii) To modify websites as they appear on the desktop — adding, removing, or modifying content. For example, the *InvisibleHand* extension adds information to shopping websites, informing the user if there's a cheaper price available on a competing retailer's website.

This paper is a study on these browser extensions and also presents a new Chrome extension-Webixnary which is a dictionary tool similar to Google's dictionary extension but with more functionalities like in-built flashcards and a whole new concept of page review-summary of a hyperlink in a given website. The paper first studies a small scale survey among 200 participants divided into 4 groups, among which, 100 are Computer Science students from the University (grouped in G1), 40 students from other branches of Engineering (grouped into G2), 40 from non-technical fields (like Law, management etc., grouped into G3), and 20 other randomly chosen participants from outside the University (in group G4). The data collected through this survey is given in Table II. The paper also investigates 60 Chrome extensions which are similar in the application we developed. Findings of this review discuss the similarities and differences among these extensions and show how our version of this application is better than the majority of the extensions which we reviewed. This paper shortly looked into the extension vulnerabilities as well and discussed how initial precautions can be taken to develop a relatively more secure extension.

II. BACKGROUND

Since the inception of the World Wide Web (1990), also known as the web, terms like surfing the Internet or browsing the Web coined up [13]. At the very same time, the first-ever web browser "WorldWideWeb" (later renamed to Nexus) was also created [56]. But web browsers didn't gain much popularity until Mosaic came up. Mosaic, also known as "the world's first most popular browser", was created in the year 1993, and 1993 became a landmark year in the browser history [20]. Then came up Netscape Navigator in 1994 with a little improvement over Mosaic, and soon Microsoft debuted Internet Explorer in 1995. A few years later Mozilla launched Firefox as Netscape Navigator faded out in 2004. Apple released its own browser Safari in 2003, and the last major breakthrough to the browser market was Google's Chrome released in 2008. The Internet world has taken a leap from there and there are now about 61 web browsers created around the globe [9], of which 18 have been discontinued. From such a big set of web browsers, they are only few which rules the browsing world of internet users, these widely used and most popular web browsers are Chrome, Firefox, Internet Explorer, Microsoft Edge for desktop and in the Mobile world (cell phones) Safari, Opera, UC are also some popular names. Among these, the current most widely used web browser is none other Google's Chrome. As of July 2017, Chrome covers 59.69% of Browser market Share Worldwide [10] "Fig.1", Table I. According to some

statistics, Google Chrome has surpassed 2 billion installs as of 2017 [24]. The next breakthrough in the world of browsing was the development of Browser Extensions [12]. With the first coming for and from Netscape browser. The development of a Browser extension is actually the creation of an extension for a specific browser and each of these browser types has their own architecture and APIs (Application Programming Interface) to create extensions. One of the original API was NPAPI (Netscape Plugin Application Programming Interface) which was first developed for Netscape browsers, beginning in the year 1995 with Netscape Navigator 2.0, this got subsequently adopted by other browsers as well. The other very popular browser at the time, Internet Explorer started supporting these extensions from version 5 released in 1999, and Firefox has supported extensions since its launch in the year 2004. Google Chrome also started supporting extensions from version 4 released in 2010. Apple's Safari has supported native extensions from its version 5 release in 2010. Finally, Microsoft Edge has supported limited extensions since March 2016. A few key points about browsers and browser extensions are:

- A web browser: it's what connects users with everything on the web.
- A browser extension: an extension to a browser is what CSS and JS are to an HTML code.

This paper mainly studies browser extensions and a brief history of web browsers was given in the previous paragraphs because browsers are the place where extensions are deployed. In this paper, for the implementation purpose and the in-depth study on browser extensions, chosen Chrome extensions along with the Chrome browser has been chosen because of the immense popularity and usage of Chrome browser [10, 11]. Due to time constraint and other limitations, we have not discussed many other browsers or their extensions. The statistical data collected from online sources for browser statistics, Chrome usage statistics, and Chrome extension statistics to support the study are presented in this paper. We also conducted a small-scale survey involving 200 participants for some experimental analysis to get similar statistics. This paper also studies the significance of these browser extensions and their application and what motivated us to create our own Chrome extension for one such application. To compare our extension we studied 60 other Chrome extensions which have similar application and functionality and discussed the similarities & differences among these extensions.

Table I. Web browsers market share around the world

Web Browsers	Market Share%
Chrome	59.69
Safari	13.85
UC Browser	7.03
Firefox	5.02
Opera	3.35
Internet Explorer	3.01
No info	8.05

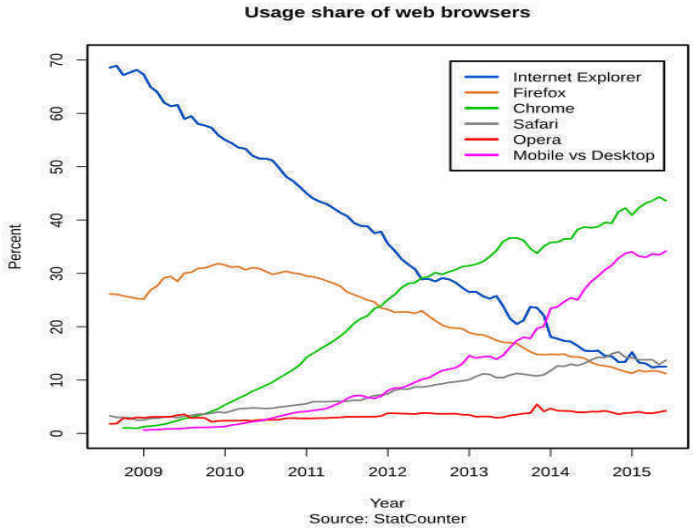


Fig. 1. Usage share of web browsers

Table II. Usage share of web browsers (only for desktop, given in %) among different groups studied

S. No.	Category	G1	G2	G3	G4
1	Participants using only one web browser	26	20	60	60
2	Participants using more than one web browser	74	80	40	40
3	Chrome as first preference	80	70	95	80
4	Firefox as first preference	16	20	5	20
5	Browser other than Chrome as first preference	20	30	5	20
6	Other browsers as second preference	62	55	40	40

III. MOTIVATION

With the increasing growth in internet usage and users (as of June 2018, 55.1% of the world's population has internet access [28]), the amount of time these users spent online/digital media has also increased, 5.9 hours a day in 2017 from 2.7 hours a day in 2008 [45]. The online activity involves a lot of different platforms like social media, YouTube, blogging websites, etc. A big part of this time online involves reading something and a lot of the reading materials or websites are in the language English. For non-native English users, it's difficult to be familiar with a lot of English words, so while reading online they also spent some non-trivial amount of time looking for the definition of the words they don't understand. What they can do to solve this issue is to look up the words in a manual dictionary or online itself. Both these process takes more time than needed and readers have to put unnecessary extra effort into doing that. So a dictionary tool in the form of a browser extension can save them that extra effort and some time. A browser extension is a piece of software that adds more functionality to a given browser. These applications are not very difficult to create. Creating an extension mainly requires knowledge of simple web technologies like HTML (Hypertext Markup Language), CSS and JavaScript. So a

browser extension was the best way to solve such problem, and also due to reasons like they are easy to create and distribute it to multiple users across nations to help readers get what they want from web more efficiently. So this paper presents Webixnary- a dictionary tool for the web. It is currently only a Chrome extension and not a cross-browser extension, by the writing of this paper.

Chrome Extension

Availability is the most important factor here. But usage and popularity matter too. As of July 2017, 59.69% people over the world use Chrome as a web browser, across multiple devices, Table I. And as a desktop browser, Chrome holds 66.43% net market share worldwide, as of October 2018 [10]. Also, Chrome has a large collection of extensions in the Chrome web store over 8500 extensions and 1,500 themes [22, 23]. And creating a Chrome extension is comparatively easier than developing other browser extensions due to a large developers' community and Chrome extension development guide provided in Chrome developers tool [1].

Also since now Chrome is also the major browser in the market, and there are now over 1 billion [24] chrome users around the world, so a larger community of users as well as developers. According to a study, statistics tell that Google has only served 70 million extensions and themes which is a mere fraction of the 2.5 billion installs by Mozilla [23]. The chrome extension users' growth is shown in "Fig.3".

IV. RELATED WORK

As per our research, we found out that this idea of a dictionary tool as a browser extension has already had some prior work done. Similar applications already exist in use. One of such extensions is from Google itself: Google Dictionary extension [44], as told we are limited in our research to only Chrome extensions, we found out that more than 50 other similar extensions are there in the Chrome web store. But most of these extensions have one or very few functionalities and compared to that, Webixnary has more features into one single tool, these extensions only share one or few of those features. Following is a detailed study on the differences and similarity between Webixnary and the other 60 Chrome extensions in that similar application idea.

For a case study, Google Dictionary Chrome extension [44] has been chosen because of its simplicity and authenticity. This extension from Google shows the definition of any selected word (double clicked or triggered keys) in a popup-bubble with a link of "More" to get more information on the word by opening a new tab and looking up the word. This definition is mostly a sentence long meaning of the word. This definition is fetched from the Google search engine itself, mostly the definition numbered as 1 or the uppermost/common most meaning of that word. The extension also provides a toolbar option on the right corner of the web page, to directly Google search any word, not necessarily has to be from the current web page a user is on; the word's definition shows up in a resized pop-up page showing the meaning from the search engine itself. The popup box displays the word's definition when it is selected shown in "Fig.2".

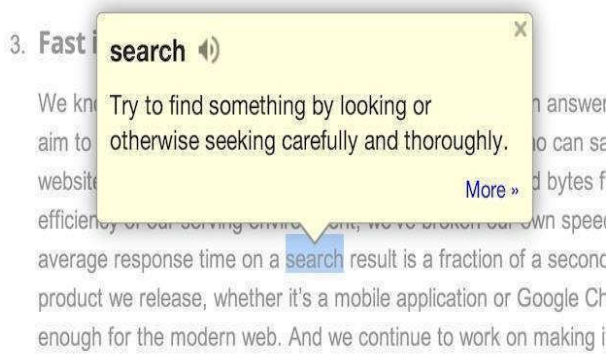


Fig. 2. Popup box displays the word's definition when it is selected.

As mentioned earlier this paper studied 60 Chrome extensions to compare it to Webixnary. Following are some of the findings:

- Majority of these extensions were only a dictionary tool with no other feature. But their presentation of the display and source of definition for the word has been from a variety of sources. For example, one extension [English Popup Dictionary] has derived its definition from Google search engine only, filtered it to top 3 meanings. Another one [Dictionary] fetched the meaning as well other things like synonyms with word from *Wiktionary* (the Wikipedia dictionary), one extension derives the definition from the Oxford Dictionary online archive. Example: [Wordbook and Oxford Dictionary Extension 0.1.29]
- Another very common feature in a lot of these extensions is the translate feature. Many extensions provided a variety of languages to choose from for translation. Example: [ImTranslator: Translator: Translator, Dictionary, TTS 14.17].
- Some of the extensions only had the translate features with no word definitions or any other feature. Example: [Dictionary 3.3.0]
- A lot of extensions provide a repository option, like saved history of the searched words which could be accessed through the *Options.* Or a downloadable file like *csv* or word file. Example: [WordBuddy 1.0], [Zinger 0.3]
- Some of these extensions have Flashcards too, which give a "word-of-the-day" based on the frequency setting of the extension or the user's choice. This flashcard could either generate random words and their definitions for users to learn new words or can fetch one of the previously looked up words. Example: [Cabulary 1.6.0], [Word Of The Day! 1.5.1]
- Some of the extensions in that set are only Flashcard extensions i.e. they don't have any dictionary tool but they generate random words/idioms sometimes for the word-of-the-day. These extensions can help assist the only dictionary base extensions like Google dictionary as flash those words looked up through it. Example*:* [Idioms and Phrases 0.4], [words. 0.33].
- Few extensions were only an archive, that is, they have no dictionary or word lookup tool or translate tool, and they just store the selected words which can

later be seen by the user along with its definition. Example: [WordBuddy 1.0]

- While a majority of these extensions are available in the Chrome web store for free. Very few were with a little fee to pay to use them. Example: [Instant Dictionary by GoogWordGuide.com 3.71].
- Some of the extensions were totally different from any of the above-discussed features. Some of them targeted to do a specific task, for instance, replace the selected word for a more difficult word. Example: [Smart Words 1.0].

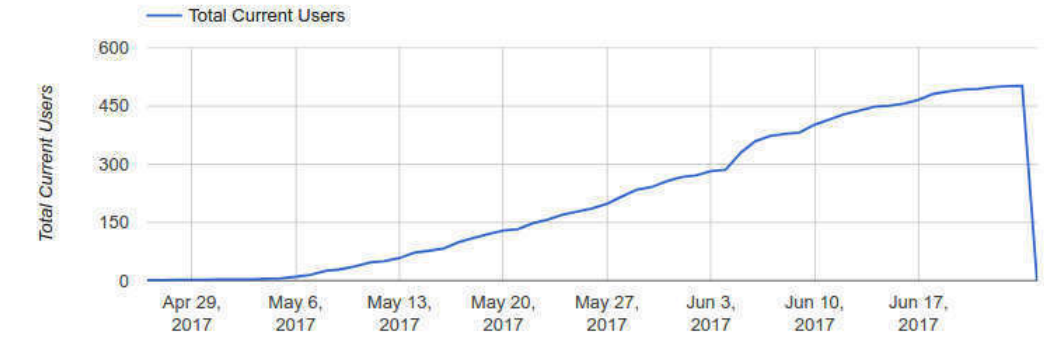


Fig. 3. Chrome extension users' growth

- Some of them worked even in offline mode and on the local system applications like PDF reader, MS Word etc. Some were just a storage-based where user selects a word and it gets saved a file. Example: [WordComplex (English Dictionary) 1.1.4].
- A lot of these extensions have a mix of above-mentioned features. Example: [Dictozo 2.0.3], [Dictionary+ 0.1.4].

From the data above, it could be inferred that a lot of prior work exists in this area. But this extension- Webixnary is an "all-in-one" package so far. The most distinguishable feature of this extension is the Page Review feature. None of the 60 extensions which have been studied has this functionality. As per the research and to best of our knowledge there's no such browser extension (limited to Chrome) with dictionary tool as well as a page review functionality. These two features along with few others can let a user look for any word on the web including hyperlinks on the go without having to open any other tab, thus the name- Webixnary (Web-Dictionary). Another advantage with this extension is, it is developed by taking care of the basic guidelines of how to avoid and/or reduce the vulnerabilities of browser extensions especially Chrome. The implementation and development of the extension followed the methods and rules given on the Chrome's developer extensions security page [47, 49, and 48].

- One such rule is: Not to inject JavaScript code inside an HTML page of an extension.
- Another important rule: Never to use HTTP (Hypertext Transfer Protocol). An HTTP connection should be avoided when requesting or sending data and an HTTPS should be preferred over it, as it has built-in security bypassing for most man-in-the-middle attacks.
- Request minimal permissions: Limiting an extensions privileges limits what a potential attacker can exploit.

The extension has been developed according to the above guidelines and thus it's relatively safer than many other extensions studied in the survey. For example, when looking at the source code of few extensions from the survey list, some of them have not taken care of the rules and methods of avoiding the extension vulnerabilities, thus making their extension more vulnerable and compromising user's privacy.

Also, through a study of 100 Chrome extensions, this data "*27 out of 100 browser extensions contain one or more vulnerabilities, for a total of 51 vulnerabilities. Among these, a set of vulnerable extensions includes 7 extensions with more than 300,000 users each*" [50], this suggests that a developer should take minimum steps to ensure the security of the extensions they are developing. Even Webixnary is not perfect, it still might have chances of being the gateway of vulnerabilities for the attackers. But as part of the future work, we will try to make it more secure and enhanced. The review of 100 Chrome Extensions for potential vulnerabilities shown in "Fig.4."

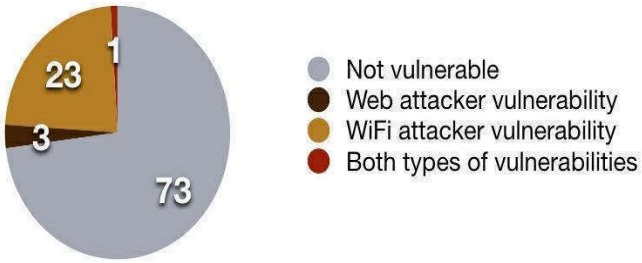


Fig. 4. Review of 100 Chrome Extensions for potential vulnerabilities

V. IMPLEMENTATION

The browser extensions add functionality to a web browser or alter and make changes to a website or a webpage. So as a proof-of-concept, this paper presents

Webixnary- a dictionary tool for internet users through a Chrome extension.

- Hardware used: HP Pavilion laptop.
- Chrome Browser version: 71.0.3578.30 (Official Build) beta (64-bit).
- Tools: HTML, CSS, JavaScript.

Chrome developers tool already has the documentation and guidelines for how to develop Chrome extensions so we have not covered the development of this extension in the paper, in depth. But a general architecture/perspective is given below:

An extension needs to:

1. Modify (read and alter) the Document Object Model (DOM) of a web page.
2. Make HTTP requests - so that the data it collects can be written to a database.
3. Store cross-domain state.
4. Read Uniform Resource Locator (URL) - so GET parameters can be passed to the extension.

A. Chrome extension architecture

Chrome extensions are built on simple web technologies like HTML, JavaScript, etc. These files are zipped into a single .crx package that a user downloads to install the extension. These files also have access to the APIs that a browser give to web pages for actions such as Http/Https Requests to HTML5 features. Following are a list of files, a chrome extension generally contain:

1. A must-have "manifest.json" file.
2. One or more HTML files (for UI).
3. A CSS(Cascading Style Sheets) file for extending the HTML files.
4. A JavaScript file or any other file (JQuery, etc.) to extend functionality to all the other extension files.

A Chrome extension is categorized as following based on their action-type:

- Browser Action
  - These stay in the toolbar.
  - They are active all the time.
  - They are accessible at all times.
- Page Action
  - They stay in the toolbar but greyed out.
  - Accessible and active only on certain pages.
- Neither browser or page action
  - They run in the background.

Among these action type, most of the Chrome extensions have a background action page which holds the main logic and state of the extension. They also often also have content scripts which can interact with web pages. In the package files, content scripts and background page communicate to each other by message passing (through Chrome APIs). Extensions can also save data through the HTML file using storage APIs like Chrome's local storage.

Manifest Files: As mentioned, every Chrome extension must have a manifest.json file. This manifest file contains information about the extension such as extension's name, version, description, permissions, content scripts etc.

Content Scripts: Content scripts are JavaScript files that are injected to a given web page. They run in the context of web pages and can directly access the Document Object Model (DOM) of these web pages, thus able to alter or override an entire web page.

Page Actions: These are the extension icons that show up next to the Omni bar. They work in two ways: with a popup or without a popup and mostly serve as the UI (user interface).

Background page: In a Chrome extension, a background page is a key component to the extension's functionality. Some key responsibilities of a background page are as following:

- It runs in the background and listens to an event.
- It is loaded when needed and unloaded when goes.
- Interact with the content scripts through message passing.
- Interact with other files like the popup page to perform any action or create an event.

B. Extensions implementation for users

Once the user installs the extension from the Chrome web store, and turn it on by clicking the icon, the extension starts running, and as it is browser action based it's active on any website, any webpage. Now to look up a word's definition, selecting or double-clicking (manually configurable options, "Fig.7") a word, will show up a pop-up box with the foremost definition from the Google search engine or any other online dictionary source (For e.g. Wiktionary or Webixcon.org), "Fig.6". A user can also search any word through the context menu option as shown in "Fig.5". Through the context menu, users can also save the word and create their own archive of words. There will also be a flashcard feature which user can select to get words fetched from either their collection or random words from any online dictionary source.

Page review: In any given web page, there's always a possibility of having some words as hyperlinks (referred to as hypertexts), some of the extensions studied in this research could fetch the definitions of these hypertexts through their context menu option. But Webixnary is not limited to just the hyperlink-word's definition, its page review option shows a popup box whenever hovered over a hypertext, the popup shows a brief content of the web page that the hyperlink is for. The brief content can be done in two ways, one way is just simply open a resized tab with that page's content, by not leaving the current web page. The second approach is better, it extracts the first few lines from top-most paragraphs and displays it like a summary in a popup bubble. Thus, the reader can get immediate insight on what the link contains and if it's worth reading or visiting further. Currently, Webixnary uses the first approach, but as future work, we'll implement the second approach.

This page review synonymously termed as the summary content of a hyperlink word. But upon further researching, Wikimedia's page review [43] came into light. So this functionality is similar to that and to some extent trying to achieve a similar objective. With the help of this feature, any web page on the web can now show the summary-page review of their hyperlink words just by hovering over the link or using some triggered key.



Fig. 5. Webixnary as context menu

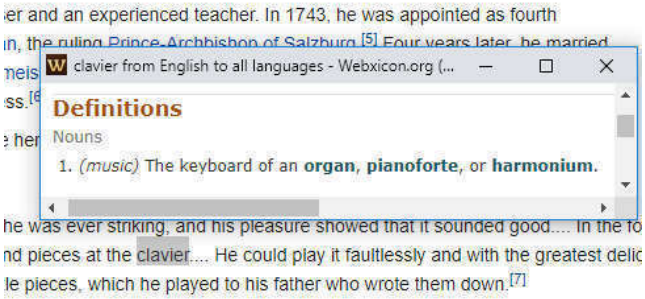


Fig. 6. Popup box with the definition of the selected word

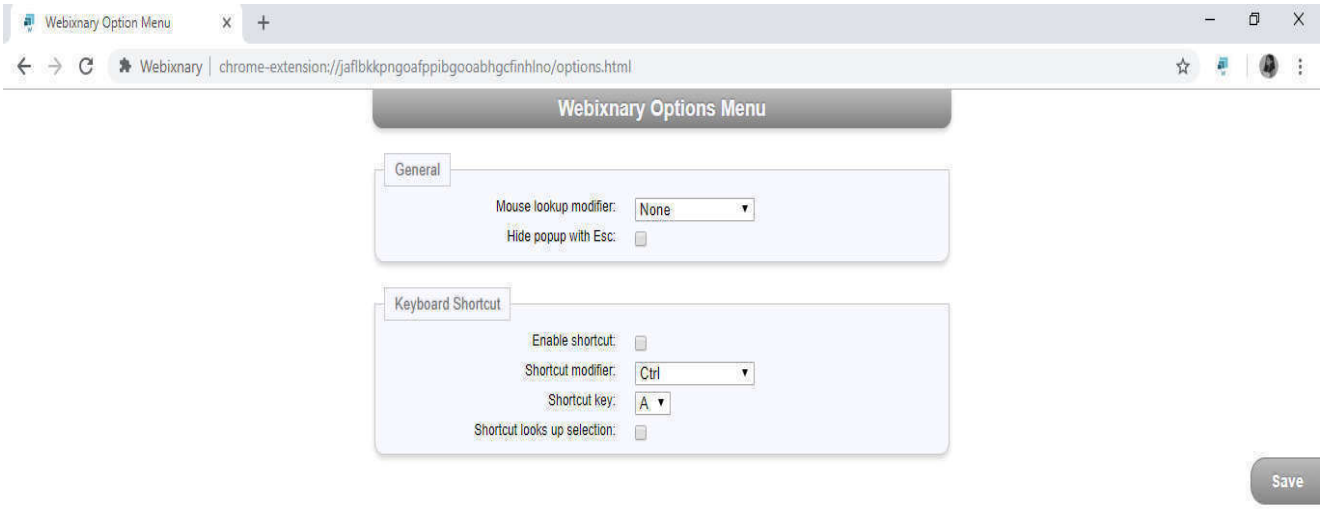


Fig. 7. Webixnary Options Menu

VI. Result and Discussion

Using simple tools like HTML, CSS and JavaScript, we created Webixnary- a dictionary a page review tool along with several other features like: translation of the word into a given language, saving the searched words (history archive) as both online vocabulary store and as downloadable csv file., text-to-speech delivery of the word in few languages, flashcards to get the word-of-the-day from either the history collection or randomly generated words. We also studied that there already have been similar work done in this area, but most of those works don't include all these functionalities (features) in one application. In addition to that, the page review feature added which have not been done previously in any such dictionary related tool as to best of our knowledge and those 60 Chrome extensions we surveyed verifies the hypothesis. We also studied the vulnerability of these browser extensions and took into consideration the essential steps required to develop a relatively more secure extension, so not to let this extension becomes a source of vulnerability for the system. This paper also conducted a small survey to collect data on the usage of web browsers among different groups (participants were mostly college students but from different field of study). The findings of the survey are shown in Table 2, which conclude that Chrome is the most popular browser in use among these participants, followed by Firefox. Not explicitly shown in the table, but a fair number of Tor browser users were also found. Also, an average of 78% of these participants uses one or more browser extensions. The survey also showed that out of 200 participants only 1% (2 participants) knew how to develop a browser extension which is a very less number given that the most of the participants were from Computer Science background and knew all the basic tools to make any browser extension. Webixnary as a context-menu option is shown in "Fig.5". Displaying the definitions of a selected word (fetched from an online dictionary source: Webxicon.org) is shown in "Fig.6". The options page of Webixnary is shown in "Fig.7".

A. Browser extension vulnerabilities

Apart from the significance of browser extensions, their development and application, there are various studies and research going on how to deal and overcome with the vulnerability of these browser extensions. Across the globe, at multiple locations, prestigious academic institutions (from IIT Delhi (India), to Universities like Stanford, MIT, CMU (USA) and even corporate platforms have done/conducted a lot of research in the safety/security and vulnerability issue of these browser extensions [31], [39], [54]). Some of them are:

- Chrome Extensions: Threat Analysis and Countermeasures [32].
- Spying Browser Extensions: Analysis and Detection [31].

- Protecting Browsers from Extension Vulnerabilities [54].
- Watching Them Watching Me: Browser. Extensions: Impact on User Privacy Awareness and Concern [34]
- Botnet in the Browser: Understanding Threats Caused by Malicious Browser Extensions [35].
- Google Chrome Browser Extensions: Another Malicious Portal [38].

These research holds a lot of significance because they directly deal with the privacy rights of any internet user. As much a browser extension adds functionality to any browser, it also adds a chance of becoming a gateway for attackers, malicious software programs. In fact, a lot of these extensions are a threat in themselves. They collect user's data like passwords, card details and other personal details with or without their permission and then exploit their privileges to interact with malicious programs websites and compromise user's privacy. They can also potentially harm other websites, other online databases or archives, steal their data and then trade those data to a third party for profit or anything in exchange.

CONCLUSION

This paper started with a short study on the evolution of web browsers followed by showing the significance of browser extensions through collected statistical data and then a detailed study on popularity and usage of various web browsers through a survey conducted among 200 participants. It also investigates into matters like how many of these participants are familiar with browser extensions and to what extent and how much they know about the development process of these extensions and their security vulnerabilities. Finally, this paper presented Webixnary, a Chrome extension with following functionalities:

- A pop-up bubble to show the definition of a selected word on a given webpage.
- Translation of the word into another language.
- Text-to-Speech (TTS)
- Saving the searched words into an online archive or in a downloadable file format.
- Flashcards to review the looked up words or to learn a new word per day.
- A most significant feature being the Page review-summary of a word or any term which is a hyperlink in the given web page.

The most important future direction we envision is to extend the summary-page review functionality of the extension, make it more informative, thorough and still concise. Also by the writing of this paper, the extension is still in development and includes only the dictionary and page review tool, other features yet to be added.

REFERENCES

[1] https://developer.chrome.com/extensions
[2] https://chrome.google.com/webstore/category/extensions
[3] https://developer.chrome.com/webstore/faq
[4] https://expandedramblings.com/index.php/google-app-statistics/
[5] http://martin.majlis.cz/chrome-extension-usage-mystery/
[6] https://www.bleepingcomputer.com/news/google/google-open-sources-some-of-chromes-usage-stats/
[7] https://en.wikipedia.org/wiki/Web_browser
[8] https://en.wikipedia.org/wiki/History_of_the_web_browser
[9] https://en.wikipedia.org/wiki/List_of_web_browsers
[10] https://en.wikipedia.org/wiki/Usage_share_of_web_browsers
[11] https://en.wikipedia.org/wiki/Comparison_of_web_browsers
[12] https://en.wikipedia.org/wiki/Browser_extension
[13] https://en.wikipedia.org/wiki/Jean_Armour_Polly
[14] http://www.internetlivestats.com/internet-users/
[15] https://en.wikipedia.org/wiki/Global_Internet_usage
[16] https://en.wikipedia.org/wiki/Google_Chrome#Extensions
[17] https://en.wikipedia.org/wiki/List_of_Firefox_extensions
[18] https://en.wikipedia.org/wiki/Add-on_(Mozilla)
[19] https://en.wikipedia.org/wiki/Internet_Explorer#Extensibility
[20] https://crossbrowsertesting. automation/history-of-web-browsers/com/blog/test-
[21] http://martin.majlis.cz/chrome-extension-usage-mystery/
[22] https://blog.chromium.org/2010/12/year-of-extensions.html
[23] https://www.extremetech.com/internet/87484-85-of-firefox-users-use-add-ons-chrome-users-just-33
[24] https://expandedramblings.com/index.php/google-app-statistics/
[25] https://www.statista.com/topics/1145/internet-usage-worldwide/
[26] https://www.statista.com/topics/2157/internet-usage-in-india/
[27] https://venturebeat.com/2015/05/28/google-chrome-now-has-over-1-billion-users/
[28] https://thenextweb.com/contributors/2017/04/11/current-global-state-internet/
[29] https://howtogeek.com/169080/beginner-geek-everything-you-need-to-know-about-browser-extensions/
[30] http://static.usenix.org/event/sec10/tech/full_papers/Bandhakavi.pdf
[31] http://precog.iiitd.edu.in/Publications_files/aa-spyingextensions.pdf
[32] http://www.cse.psu.edu/~trj1/cse597-s13/docs/chrome_ext_ndss_12.pdf
[33] http://www.cse.chalmers.se/~andrei/codaspy17.pdf
[34] http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2017/09/watching-them-watching-me-browser-extensions-impact-on-user-privacy-awareness-and-concerns.pdf
[35] http://homepages.cs.ncl.ac.uk/feng.hao/files/Botnet_In_The_Browser.pdf
[36] https://pdfs.semanticscholar.org/d24e/f1f8c2c9bfbd2bd552b1ba516e4147cf6423.pdf
[37] https://cseweb.ucsd.edu/~ldekoven/publications/malicious_browser-extensions_at_scale.pdf
[38] http://www.cs.tufts.edu/comp/116/archive/fall2017/omacdougal.pdf
[39] http://crypto.stanford.edu/~jcm/papers/pwdhash.pdf
[40] https://drona.csa.iisc.ac.in/~vg/papers/acsac2009a/acsac2009a.pdf

[41] https://www.computer.org/csdl/proceedings/sp/2017/5533/00/07958618.pdf
[42] http://dongs.co/~xinshu/papers/isc12.pdf
[43] https://www.mediawiki.org/wiki/API:Main_page/en
[44] https://chrome.google.com/webstore/detail/google-dictionary-by-goog/mgijmajocgfcbeboacabfgobmjgjcoja
[45] https://in.pcmag.com/netflix/121788/tech-addiction-by-the-numbers-how-much-time-we-spend-online
[46] https://thenextweb.com/contributors/2017/04/11/current-global-state-internet/
[47] https://developer.chrome.com/extensions/security
[48] https://www.godaddy.com/garage/security-tips-plugins-themes-extensions/
[49] https://blog.chromium.org/2008/10/new-approach-to-browser-security-google.html
[50] https://www.techrepublic.com/blog/it-security/chrome-extensions-are-vulnerable-advantage-bad-guys/
[51] https://security-assessment.com/files/presentations/liverani_freeman_abusing_firefox_extensions_defcon17.pdf
[52] https://www.techrepublic.com/blog/it-security/some-firefox-extensions-may-be-exploited-to-install-malware/
[53] https://www.techrepublic.com/blog/it-security/chrome-extensions-potential-cracks-in-the-armor/
[54] https://www.adambarth.com/papers/2010/barth-felt-saxena-boodman.pdf
[55] https://dspace.mit.edu/bitstream/handle/1721.1/85404/870307642-MIT.pdf?sequence=2
[56] https://en.wikipedia.org/wiki/WorldWideWeb

GOOGLE CHROME EXTENSIONS SURVEYED

The following Chrome extensions were studied [in the related work section IV]:

[Google Dictionary (by Google) 4.0.8], [Dictionary 2.5.4], [Dictionarist – Instant Dictionary 1.2], [Instant Dictionary 0.0.6], [Instant Dictionary by GoogWordGuide.com 3.71], [English Popup Dictionary 1.4], [Click dictionary 1.3.8], [Dictionary 3.3.0], [Dictionary (dictionarynow.info) 2.5.4], [WordMaster 1.0.2], [WordComplex (English Dictionary) 1.1.4], [Words Discoverer: Expand your vocabulary 2.10.1], [Vocabulary Boost 1.0.27], [ImTranslator: Translator: Translator, Dictionary, TTS 14.17], [Cabulary 1.6.0], [Wordgram – vocabulary manager 1.0.5], [Smart Words 1.0], [Vocablurry 2.0.0.5], [Oh My Word! 1.3.10], [CravingWords 1.0.10], [WordBuddy 1.0], [Dictozo 2.0.3], [Vocabulary Builder 1.0], [eJOY English – Word Master 1.7.0], [words. 0.33], [Zinger 0.3], [WordSwap 0.0.2], [Word Of The Day! 1.5.1], [Remember It 0.3.5], [Fluentcards Dictionary 0.7.8], [Exactly: English Dictionary and Learning List 1.70], [Vocabulary in Context 1.8], [Idioms and Phrases 0.4], [English Vocabulary Card Creator 1.1.1], [BeFluent – For English phrases and sentences 0.4.14], [Dictionary+ 0.1.4], [QwickUp 0.11], [InstaWord 2.0], [A Dict 1.0.6], [Dicty 1.0.6], [Dictionary Lookup 4.3.4], [Vocabulary Expander 2.4.1], [English Dictionary Lookup 1.1.2.0], [Wordbook and Oxford Dictionary Extension 0.1.29], [Vocabla: translate English words 2.2.2], [Urban Dictionary Instant word lookup 0.0.9], [My Dictionary 1.3.1], [Dev Dictionary 2.1.5], [Rememberry – Translate and Memorize 0.4.8], [TheFreeDictionary 1.2], [Leo Dictionary Widget 2.4], [English vocabulary expander 2018.80], [Power Thesaurus 2.0.8], [Twinword Finder 0.2.9960], [Wikipedia Search 9.0], [SearchBar 0.11.3], [Cambridge Dictionary 1.1], [One Click Dictionary 3.0], [Define 2.0.1], [Dictionary, Thesaurus & Reference 4.1.1], [Urban Dictionary Search Tool 1.80].

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A WAY TO SECURE THE DATA IN CLOUD DATA STORAGE BY USING CLOUD DATA COMPRESSION MECHANISM

Saidhbi Sheik [#1] and Thirupathi Rao Komati[#2]
[1#]Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu,
[#2]Professor and Associate Dean-Academics, K L University, AP.
[1#] {sfajju.syed@gmail.com}
[2#] {profktrao@gmail.com}

ABSTRACT

Cloud computing significantly plays a role in the aspect of effective resource utilization and service consumption. Irrespective of the type of clouds (ex. Private, public, hybrid or inter-cloud), every service providers concentrates on the data residing in cloud servers. Each and every moment, the researchers and scholars are proposing multiplicity of security algorithms to secure cloud data during the transactions. Most of the cloud data secure algorithms are focusing on the way to secure to cloud data in a single direction by using cryptographic algorithms. In this research paper focuses on a new direction to combine the features of data compression with the cloud data in order to secure the cloud data storage.

Key words: Cloud, data, storage and compression.

1.INTRODUCTION

Either to consider a Profitable or non-profitable organizations, dedicated resource utilization brings more elevation in the economic impact and makes huge loss. In order to overcome this defect, every clients hunt for a new technology to solve their demand with minimum effort. In this aspect, the cloud computing provides an excellent environment for the resource seekers over the network. Most of the categories, the cloud service providers are not consider the secure way of data transaction under the public cloud. But, in the same time, private cloud give more attention to secure the data resides in its cloud servers as well as to maintain the security for confidential cloud data. The general cloud storage mechanism is comprised with two major components such as data and its applications. Both data and applications are always handling with the help of cloud data owner and cloud service provider [5].

The most challenging task in the cloud data servers are focusing on the handling of residing data under the category of private and confidential sector classification. In order to ensure the secure data in cloud storage by using a cryptographic mechanism apply encryption on the storage sector and decryption on the authenticated receiving sector [1]. In the aspect of applying cryptographic algorithms such as RSA (Rivest, Shamir and Aldimer) and AES (Advanced Encryption Standard) regarding to secure cloud data using its own way of number theoretical information's and key exchange values [2]. The data or information is required to store in cloud service provider to keep as it is not guarantee for secure data. Because of this reason, the researchers concentrate on cryptographic mechanism.

It provides only the crypto mechanism as the solution provider in order to secure the cloud data residing in Cloud

ISBN 978-81-939386-1-4

Service providers. Most of the cloud security algorithms are focusing on the above said algorithms with performance analysis and comparisons of its required amount of storage [3]. The following block diagram (figure 1) depicted in the existing security mechanism in cloud data. In this research paper to frame, a challenging architectural framework as well as to implement the secure cloud data with data compression mechanism.

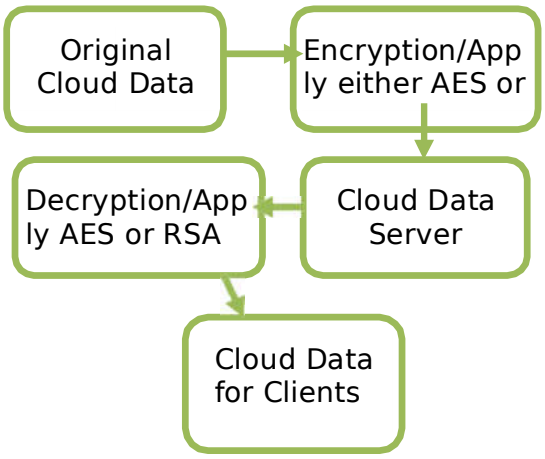


Figure 1. Existing Cloud data secure mechanism

Data compression is used in most of the cases to reduce the original size of the data or information without affect its originality as well as the number of bits. Most of the cases , the compression algorithms are classified in to two major categories : Lossless and Lossy . The Lossy algorithm is indicates few data bits are loss during the retrieval or reproducing environment. At the same time, the Lossless never makes such kind of bits or data loss during the retrieval environment [4].

II. LITERATURE REVIEW

Huffman Coding Huffman coding [7] is an entropy encoding algorithm used for lossless data compression. It uses a specific method for choosing the representation for each symbol, resulting in a prefix-free code that expresses the most common characters using shorter strings of bits than are used for less common source symbols. Huffman coding is optimal when the probability of each input symbol is a negative power of two. Prefix-free codes tend to have slight inefficiency on small alphabets, where probabilities often fall between these optimal points. "Blocking", or expanding the alphabet size by coalescing multiple symbols into "words" of fixed or variable-length before Huffman coding, usually helps, especially when adjacent symbols are correlated. Prediction by Partial Matching (PPM) [8,9] is an adaptive statistical data compression technique based on context modeling and prediction. In general, PPM predicts the probability of a given character based on a given number of characters that immediately precede it. Predictions are usually reduced to symbol rankings. The number of previous symbols, n, determines the order of the PPM model which is denoted as PPM(n). Unbounded variants where the context has no length limitations also exist and are denoted as PPM*. If no prediction can be made based on all n context symbols a prediction is attempted with just n-1 symbols. This process is repeated until a match is found or no more symbols remain in context. At that point a fixed prediction is made. PPM is conceptually simple, but often computationally expensive. Much of the work in optimizing a PPM model is handling inputs that have not already occurred in the input stream[7]. The obvious way to handle them is to create a "neverseen" symbol which triggers the escape sequence. But what probability should be assigned to a symbol that has never been seen. This is called the zero-frequency problem. PPM compression implementations vary greatly in other details. The actual symbol selection is usually recorded using arithmetic coding, though it is also possible to use Huffman encoding or even some type of dictionary coding technique. The underlying model used in most PPM algorithms can also be extended to predict multiple symbols. The symbol size is usually static, typically a single byte, which makes generic handling of any file format easy

LZ77: LZ77 algorithms achieve compression by replacing repeated occurrences of data with references to a single copy of that data existing earlier in the input (uncompressed) data stream. A match is encoded by a pair of numbers called a length distance pair, which is equivalent to the statement "each of the next length characters is characters behind it in the uncompressed stream". To spot matches, the encoder must keep track of some amount of the most recent data, such as the last 2 kB, 4 kB, or 32 kB. The structure in which this data is held is called a sliding window, which is why LZ77 is sometimes called sliding window compression. The encoder needs to keep this data to look for matches, and the decoder needs to keep this data to interpret the matches the encoder refers to. The larger the sliding window is, the longer back the encoder may search for creating references. It is not only acceptable but frequently useful to allow length-distance pairs to specify a length that actually exceeds the distance. As a copy command, this is puzzling: "Go back four characters and copy e10characters from that position into the current position"[7]. How can ten characters be copied over when only four of them are actually in the buffer? Tackling one byte at a time, there is no problem serving this request, because as a byte is copied over, it may be fed again as input to the copy command. When the copy-from position makes it to the initial destination position, it is consequently fed data that was pasted from the beginning of the copy-from position. The operation is thus equivalent to the statement "copy the data you were given and repetitively paste it until it fits".

LZ78: LZ78algorithms achieve compression by replacing repeated occurrences of data with references to a dictionary that is built based on the input data stream. Each dictionary entry is of the form dictionary[...] = {index, character}, where index is the index to a previous dictionary entry, and character is appended to the string represented by dictionary[index]. For example, "abc" would be stored (in reverse order) as follows: dictionary[k] = {j, 'c'}, dictionary[j] = {i, 'b'} , dictionary[i] = {0, 'a'}, where an index of 0 implies the end of a string. The algorithm initializes last matching index = 0 and next available index = 1. For each character of the input stream, the dictionary is searched for a match: {last matching index, character}. If a match is found, then last matching index is set to the index of the matching entry, and nothing is output. If a match is not found, then a new dictionary entry is created: dictionary[next available index] = {last matching index, character}, and the algorithm outputs last matching index, followed by character, then resets

last matching index = 0 and increments next available index. Once the dictionary is full, no more entries are added. When the end in th of the input stream is reached, the algorithm outputs last matching index. It is very important to know that the strings stored in the dictionary is in the reversed order[11-13]. LZW is an LZ78 based algorithm that uses a dictionary pre-initialized with all possible symbols. The main improvement of LZW is that when a match is not found, the current input stream character is assumed that it will be the first character of an existing string in the dictionary (since the dictionary is initialized with all possible characters), so only the last matching index is output (which may be the pre-initialized dictionary index corresponding to the previous symbol.

III.RELATED WORK

The related work regarding to secure the data is specified in a general frame work not a specific boundary to describe the combination of data compression along with cryptographic mechanism. The following diagram (figure 2) is clearly indicates the existing work modules,

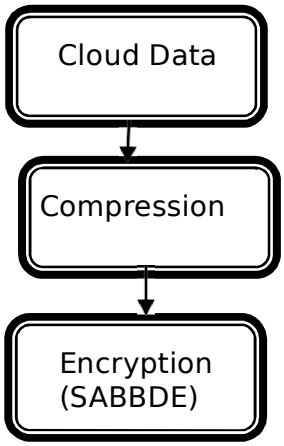


Figure 2 Existing Data security in the combination of compression and encryption

In our research work, we have proposed a novel model that combines compression algorithm and crypto mechanism. After analyzing the available compression algorithms like Arithmetic coding, Huffman coding, LZ78 and LZW, we have suggested a new compression algorithm, named as Parallelized Sparse Data Compression Algorithm (PSDCA). It offers significant advantages in saving more storage space by avoiding the wastage of storage locations by eliminating the memory entries with values of “0”. Here, as a unique provision, if "1" exceeds the threshold limit, then "0" will be stored.

Measuring Compression Performances

Typically, the performance measure proves the efficiency of a compression technique or otherwise i.e. whether it is compatible with any particular criteria. Based on the nature of the application, the criteria is selected to measure the performance of compression algorithms. In fact, time complexity and space complexity are usually regarded as the most significant criteria, used interchangeably. The compression behavior largely is determined by the category of compression algorithm chosen, whether lossy or lossless.

A major advantage of the proposed framework is that it has the potential to

reduce the accessing time between cloud clients via cloud servers, apart from less cloud storage on account of wastage by utilizing "NULL" values. However, a likely drawback could be seen in the unusually longer execution time while matching the missing storage values in the reproduction section due to slow connections in the communication channel. The comparative performance analysis is factored on the cloud storage efficiency, quality of data in retrieval efficiency and time of execution efficiency. The following chart depicts the comparative analysis between various compression algorithms such as Arithmetic Coding, Huffman Coding, LZ78, LZW and PSDCA.

Table 1
Performance Comparisons with PSDCA

Original Data			
File Name	**File Size**	**File Type**	**Parallel Arithmetic Coding**
ERR003495	1899995323	PDF	458246988
ERR227988	569784647	PNG	567704301
ERR009295	470811430	BMP	217694610
ERR092485	611354745	MP3	585467042
SRR628456	1075151912	WAV	633317808
SRR624457	603227492	DLL	153113767
SRR624558	217514321	C++	23062960
SRR624659	1608147955	Stream	774870508
SRR624850	4985746	TXT	1052673
SRR624951	5045745	TXT	1231402

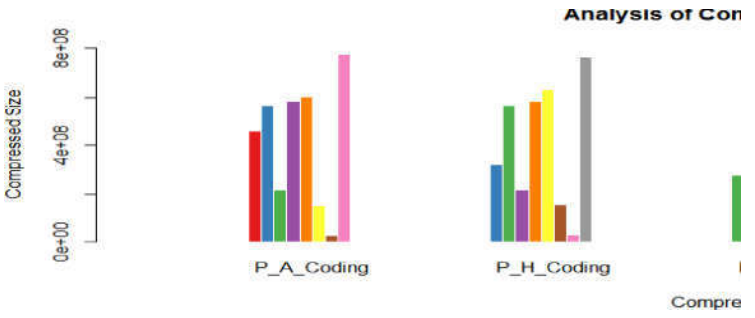


Figure 3: Analysis of compression Algorithms

The security in cloud data sector is not only the responsible for service provider alone; the owner of the data is also participating in order to store the data. Each and every IT infrastructure for cloud service provider maintains their own infrastructure in order to protect the data from unauthorized users. The data storage in the cloud servers as static category may bring a high level of security risk at this time and to create a questioner window for security treat towards its storage. According to the National Institute of Standards and Technology (NIST), the cloud models are classified into a standard way such as: Organizational Model and Service Model [6]. Most of the cases the organizational models are concentrating on the security issues related its storage infrastructures and the service model focus on the security issues related with the service utilization among the different users or clients. It will clearly depict in the following figure 4.

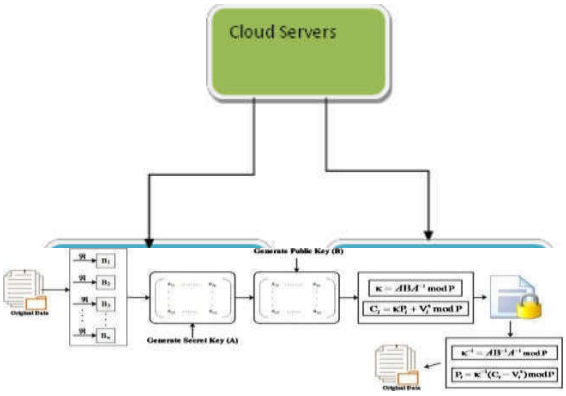


Figure 4 Cloud Service Model Classifications

Whenever to go for the cloud service storage architectural framework, it emphasis on the risk for malicious data as well as the data loss. In order to allocate the storage space for the incoming data or information, in general it assigned a proposer link with the succeeding memory locations. If the data is not a compressed one, it may require more memory space for accommodating within the existing frame and to further encryption mechanism also make a room for excess conflict in the storage infrastructure. In order to avoid such kind of storage space wastage as well as the complication access is eliminate with the help of this proposed work in this research paper.

The service model is always linked with the retrieval operation whenever it will require by the cloud service clients or users. Irrespective of the cloud type, the service related with cloud computing is always considered as software as a

Figure 5 Encryption procedure

service (SaaS) , Infrastructure as a Service (IaaS) and Platform as a Service (PaaS).

The designed architecture for cloud computing data storage is varying from vendor to vendor based on the applications. In most of the cloud service storage contents are focusing on multimedia data in its environment over the network. This drawback is overcome by this proposed research work publication.

No. of Attributes	Security Performance (%)				
	KP-ABE	CP-ABE	CP-ABE-WP	EPPDR	SABBDE
2	59.27	58.97	58.50	62.99	68.99
4	66.75	66.52	67.00	63.74	69.74
6	70.90	73.17	73.25	69.92	78.92
8	78.44	76.62	78.45	76.53	82.53
10	75.19	76.91	76.97	79.14	87.14
12	83.93	83.19	83.41	83.42	92.42
14	83.96	83.26	83.18	85.91	92.91

IV. PROPOSED WORK

The data for the transmission is used for the encryption by using Secured Adaptive Block Based Data Encryption is combined with data compression algorithm before it will store into cloud storage. In Adaptive Block Based Data Encryption each block of data is encrypted using a random number. i.e use only one random number at beginning of encryption and that can used recursively to all the data blocks using one-way hash function. This random number must be shared securely between the participants before the encryption by a one pass method.

In order to eliminate the threat for data loss is managed by the implementation of an exact link between the segmented storage element in the cloud server as well as the service providers. In most of the cases, the data retrieval or decoding information face a problem at the

ISBN 978-81-939386-1-4

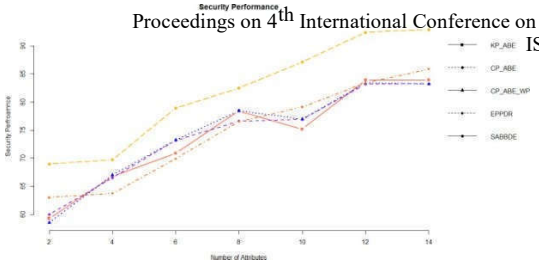


Figure 6: Analysis of different security algorithms

situation of inactive intermediate cloud servers. Such demerit or drawback is handled with the help of multiple copy of (replication) compressed as well as encrypted data to make it available in redundancy mechanism.

The cloud data which one is ready for transmission is encrypted along with compression in default manner and it will transfer towards the cloud server storage location.

Table 2:Security performance analysis for maximum number of attributes

The number of key attributes and the file sizes are the major factors in encryption and decryption process. The time requirement is more for large file size and even number of attributes increased. Let us discuss the performance analysis of time, security performance and speed for each method

Table 3: Encryption time of different files sizes and different algorithms

File Size	Encryption Time (ms)				
	KP-ABE	CP-ABE	CP-ABE-WP	EPPDR	SABBDE
10MB	93.67	92.45	90.23	89.32	65.89
100MB	189.23	183.38	178.83	172.35	142.44
1GB	484.26	436.45	418.28	389.24	338.49
2GB	784.44	764.87	755.88	684.43	656.29
3GB	1024.56	1113.34	1023.56	968.75	924.56
5GB	2346.58	2289.52	2212.87	1759.35	1649.70

In most cases, the performance of security algorithms are assessed with the help of its run-time efficiency and it will not focus on its storage. In this implementation, it gives an equal heaviness for both run time as well as storage efficiency for the cloud data residing in the cloud data server. The above graph (figure 6) clearly depicted the detailed performance evaluation.

V. CONCLUSION

Security is mostly concerned in each and every data transactions over the internet among the different users. In the same aspect, the researchers are struggling to propose a standard architectural framework in order to save the cloud data residing in cloud service provider with different mechanism. In this research paper provide an excellent roadmap to ensure the cloud data security with the help of compression and encryption mechanism.

REF

REFERENCES

[1]. " Securing and Managing Data for Cloud Storage applications using High Throughput Compression (HTC) ", International Journal of Advanced Scientific and Technical Research , Issue 5 volume 3, May-June ,2015 , ISSN 2249-9954.

[2]. "Developing Secure Cloud Storage System by Applying AES and RSA cryptography algorithms with Role based Access Control Model", International Journal of Computer Applications (ISSN: 0975-8887), Volume 118-No 12, May 2015.

[3]. "Secure User Data in Cloud Computing Using Encryption Algorithms" by Rachna Arora, Anshu Parashar , International Journal of Engineering Research and Applications , ISSN: 2248-9622, Vol-3, Issue 4, Jul-Aug 2013,pp.1922-1926.

[4]. "Data Security Using Compression and Cryptography Techniques" by Ruchita Sharma and Swarnalata Bollavarapu, International Journal of Computer Applications (ISSN: 0975-8887), Volume 117-No 14, May 2015.

[5] "Data Security Algorithms for Cloud Storage System using Cryptographic Method" by Prakash G L, Dr. Manish Prateek, and Dr. Inder Singh, International Journal of Scientific & Engineering Research, Volume 5, Issue 3, March -2014, ISSN 2229-5518.

[6]. "New mechanism for Cloud Computing Storage Security" by Almokhtar Ait El Mrabti, Najim Ammari, Anas Abou El Kalam, Abdellah Ait Ouahman OSCARS Laboratory, National School of Applied Sciences,Cadi Ayyad University ,arrakesh, Morocco,Mina De Montfort ,ARTIMIA, 75 Street Guy Mˆoquet, 92240 Malakoff, France , (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016.

[7]C:\Documents and Settings\DELL\Local Settings\temp\IM\LZSS (LZ77) Discussion and Implementation.mht.

[8] D.A. Huffman, "A Method for the Construction of Minimum Redundancy Codes", Proceedings of the I.R.E., September 1952, pp 1098-1102.

[9] T. Bell, J. Cleary, and I. Witten, "Data compression using adaptive coding and partial string matching," IEEE Transactions on Communications, Vol. 32 (4), p. 396-402, 1984.

[10] A. Moffat, Implementing the PPM data compression scheme, IEEE Transactions on Communications,Vol. 38 (11), pp. 1917-1921,November 1990**.**

[11] Ziv, J., & Lempel, A. "A Universal Algorithm for Sequential Data Compression," IEEE Transactions on Information Theory, 23(3), pp.337-343, May 1977.

[12] Ziv, J., & Lempel, A. "Compression of individual sequences via variable-rate coding," IEEE Trans. Inform. Theory, 24(5), 530-536, September 1978.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A Survey on Face Recognition Techniques and Approaches

Nikita Mor
Department of Computer Science Engineering, Faculty of Science and Technology, The ICFAI Foundation for Higher Education, Hyderabad, India
nikitamor116@gmail.com

S. Vairachilai
Department of Computer Science Engineering, Faculty of Science and Technology, The ICFAI Foundation for Higher Education, Hyderabad, India
vairachilai2676@gmail.com

Brahma Naidu Kakarla
Department of Computer Science Engineering, Faculty of Science and Technology, The ICFAI Foundation for Higher Education, Hyderabad, India
brahmanaidu@ifheindia.org

***Abstract*— Biometrics is a framework which is used to recognize human based on their physical and social characteristics. Face acknowledgement is a dynamic research zone in areas of face detection, security, and personal computer (PC) vision. Each face has its own uniqueness that portrays the personality of human. Because of its uniqueness it can be utilized for verification and access control in various applications. In this paper, the face acknowledgment and facial element is a key zone of examination. This paper incorporates the study of various face acknowledgement strategy and techniques which are professed to give a viable and exact technique for face acknowledgment. Another model for perceiving face is additionally presented in this paper. This model is executed in the imminent future and the execution is contrasted and compared with other methodology.**

***Keywords*— CNN, Feature Extraction, Deep Learning, Face Recognition, LDA, Face Images.**

I. INTRODUCTION

Deep learning gives a characteristic method to acquire portrayals from information without depending on created descriptors. Biometric security frameworks depend on facial attributes confront a testing assignment because of changeability in the intrapersonal facial appearance of subjects followed to components, such as light, demeanor, and maturing. People frequently utilize countenances to perceive people and progressions in registering ability in the course of recent decades, which currently empowers comparable acknowledgments consequently [1]. People can perceive faces effectively and easily, yet in the territory of picture investigation and PC vision it stayed as a troublesome issue on which numerous long periods of research is going on. It is valuable to have a machine perform design acknowledgment. Specifically, machines which can peruse confront pictures are extremely financially savvy. The space of machine learning and information investigation is often utilized in different applications. Some of them are utilized for decreasing the measure of information mid insight framework advancements and also utilized for specifically delivering the results to the applications. Among them the acknowledgment and order is a traditional issue of machine learning. In human face data preparing, human face location is the most vital process in applications such as video observation, confront acknowledgment, and PC helped fabricating. In any case, because of the varieties in foundation, light, posture, and outward appearances, confront recognition is yet an incredible test [2,3]. This paper audits the face acknowledgment framework and ongoing work done in this field. The proposed idea is to discover the answer for face acknowledgment utilizing neural system-based applications. In the process of face identification approach, the proposed work will execute confront acknowledgment strategy utilizing deep inclining of neural system procedures.

II. BACKGROUND

The foundation of an investigation is an imperative piece of an examination paper. It gives the unique circumstance and reason for the examination. Consequently, there is requirement for foundation ponder that adds to get ready the proposed framework.

A. Face Recognition

Face acknowledgment is turning into a functioning examination territory spreading over a few trains, for example, picture handling, design acknowledgment, PC vision, neural systems, subjective science, neuroscience, brain size, and, physiology. It is a devoted procedure, not simply a use of the general protest acknowledgment process. It is additionally the portrayal of the most marvelous limits of human vision. Face acknowledgment has been a standout among the most fascinating and essential research fields in the previous two decades. The reasons originate from the need of programmed acknowledgments and observation frameworks, the enthusiasm for human visual framework on face

acknowledgment, and the plan of human-PC interface, and so on [4]. Face acknowledgment portrays a biometric innovation that endeavors to set up a person's personality, otherwise called facial acknowledgment or face discovery. The procedure works utilizing a PC application that catches an advanced picture of a person's face (some of the time taken from a video casing) and looks at it to pictures in a database of put away records [5].

B. *Face Recognition Structure*

The structure of face recognition systems is shown in Figure 1 comprising of three noteworthy advances: (i) Gaining of Face Data (ii) Extracting Face Feature (iii) Recognition of Face.

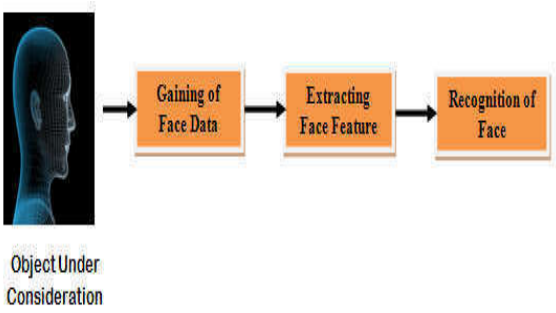


Fig. 1. Face Recognition Systems [6]

(i) Gaining of Face Data

The initial phase in the face acknowledgment framework is securing and processing of face data. In this progression, confront pictures is gathered from various sources. The sources might be camera or promptly accessible face picture database on the site. The gathered face pictures ought to have the posture, brightening, appearance and variety with the end goal to check the execution of the face acknowledgment framework under these conditions. Preparing of face database causes effect on the execution of face acknowledgment frameworks because of the changes in the brightening condition, foundation, lighting conditions and camera separation. In this manner input picture is standardized and some preprocessing techniques are applied on the information picture [7].

(ii) Extracting Face Feature

Highlight extraction process can be characterized as the way towards extricating pertinent data from a face picture. In this process a scientific portrayal of unique picture called a biometric layout is created that will shape the premise (vector) of any acknowledgment undertaking. Later, these separated highlights are utilized in acknowledgment. A grayscale pixel is considered as introductory element.

(iii) Recognition of Face

When the highlights are extricated and chosen, the following stage is to group the picture. Appearance-based face acknowledgment calculations utilize a wide assortment of grouping techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA), in order to compare between appearances of a similar individual and diverse people after all the face pictures in database are checked with important highlights. In most cases, highlight extraction and acknowledgment process is done at the same time.

C. *Linear Discriminant Analysis (LDA)*

LDA is generally used to generate linear blends of highlights. In contrast to PCA, LDA is used to demonstrate the contrasting features between classes. LDA performs well when it considers just two classes. In particular, it requires information about various classes which are farther away from one another, and also similar classes that are close to one another. Subsequently, LDA acquires differenced projection vectors for each class. Multi-class LDA calculations are done when there are more than two classes that are utilized. Linear Discriminant Analysis (LDA) is most generally utilized as dimensionality decrease strategy in the pre-handling situations for example characterization and machine learning applications. The objective is to extend a dataset onto a lower-dimensional space with great class-detachability capability which are kept together to keep away from over fitting and also lessen computational expenses [8].

D. *Diverse Approaches of LDA*

Informational collections can be changed and test vectors can be arranged in a different space by two distinct methodologies [9].

(i) Class-subordinate change

This sort of methodology is used to maximize the proportion of between-class fluctuation to within-class difference. The principle objective is to amplify this proportion so that sufficient class distinguishableness is achieved. The class-particular compose approach includes utilizing two advancing criteria for changing the informational collections freely.

(ii) Class-autonomous change

This methodology includes augmenting the proportion of general fluctuation to inside class difference. This methodology utilizes a single improving rule to change the informational indexes and therefore all the information independent of their

class character are changed. In this kind of LDA, each class is considered as a different class.

III. LITERATURE SURVEY

The different sorts of face discovery procedures are useful in writing some key commitments which are offering help in the outline and improvement of the precise face recognition. A portion of the basic articles are incorporated into this segment. Wael AbdAlmageed et al. [10] presented the technique and the framework for face acknowledgment utilizing different Deep learning models. In this portrayal, a face picture is handled by a few postures particularly by Deep Convolutional Neural System (CNN) models to create various posture particular highlights. 3-Dimensional (3D) rendering is utilized to produce numerous faces from the information picture. Affectability of the acknowledgment framework to present varieties is diminished since creators utilize an outfit of posture particular CNN highlights. Creators presented broad trial results on the impact of milestone identification. CNN postures show determination on the execution of the acknowledgment pipeline. This portrayal accomplishes preferable outcomes over the best in class on Intelligence Advanced Research Projects Activity (IARPA's CS2) and National Institute of Standards and Technology's IARPA Janus Benchmark (NIST's IJB) in both confirmation and recognizable proof (i.e. look) errands. Changxing Ding et al. [11] proposed a novel face recognizable proof system equipped for taking care of the full scope of posture varieties inside ±90° of yaw. The proposed structure initially changes the first posture invariant face acknowledgment issue into an incomplete frontal face acknowledgment issue. A strong fix-based face portrayal conspire is then created to speak to the incorporated fractional frontal countenances. For each fix, a different word reference is learnt, under the proposed system it performs multiple tasks learning plan. The different word reference changes the highlights of various postures into a discriminative subspace. lastly, confront coordination is performed at fix level instead of at all-encompassing level. Broad and precise experimentation on Facial Recognition Technology (FERET), Carnegie Mellon University Illumination and Expression (CMU-PIE), and Multi-PIE databases demonstrates that the proposed technique reliably outflanks single-assignment based baselines and in addition results in the best in class strategies for the posture issue. Creators additionally broaden the proposed calculation for the unconstrained face confirmation issue and accomplish top-level execution on the testing Labeled Faces in the Wild (LFW) informational collection. Deep learning gives a characteristic method to acquire portrayals from information without depending close by made descriptors. Xue-wen Chen et al. [12] proposed to learn Deep component portrayals utilizing unsupervised and administered learning in a fell mold to deliver conventionally enlightening yet class particular highlights. The proposed technique can take full preferred standpoint of the accessibility of substantial scale unlabeled information and learn discriminative highlights from nonexclusive highlights. It is then connected to numerous basic facial areas to acquire multi-channel Deep facial portrayals for face acknowledgment. The viability of the proposed highlight portrayals is approved on both controlled (i.e., broadened Yale-B, Yale, and AR) and uncontrolled benchmark confront databases. Exploratory outcomes demonstrate its adequacy. Deep convolutional neural systems have as of late demonstrated to a great degree viable for troublesome face acknowledgment issues in uncontrolled settings. To prepare such systems, huge preparing sets are required with a large number of named pictures. In the work, proposed by Guosheng Hu et al. [13] a technique to produce extensive preparing datasets of manufactured pictures by compositing genuine face pictures in a given dataset. Creators demonstrate that this technique empowers to take in models from as few as 10,000 preparing pictures, which perform keeping pace with models prepared from 500,000 pictures. Utilizing this methodology, they additionally get cutting edge results on the CASIA NIR-VIS2.0 heterogeneous face acknowledgment dataset. Biometric security frameworks dependent on facial attributes confront a testing assignment because of changeability in the intrapersonal facial appearance of subjects followed to elements, for example, present, light, demeanor and maturing. Hachim El Khiyari et al. [14] proposed a Deep learning and set-based way to deal with face acknowledgment subject to maturing. The pictures for each subject taken at different occasions are treated as a solitary set or, in other words the sets of pictures having a place with different subjects. Facial highlights are separated utilizing a convolution neural system for Deep learning. This trial result demonstrated that set-based acknowledgment performs superior to anything the singleton-based methodology for both face distinguishing proof and face check.

IV. PROPOSED WORK

In the proposed work the key capacities and techniques are featured which are utilized to process the face picture information. Their working procedure is portrayed as the given model, the framework acknowledges three face models for preparing reason. Among them first posture is taken from front side,

second is taken from left side and the third face picture is taken from right side. All the face pictures are handled in next stage for bi-separating, these pictures are changed into six incomplete stages. After transformation of these countenances into six sections, the arrangement is made to characterize the picture classes. These picture classes are utilized with the LDA which includes extraction calculation. The LDA highlights and the class definition are safeguarded in a database. Then a neural system is set up for preparing with the separated highlights. In this way the client gives the setting parameters to the neural system., utilizing these parameters framework performs preparing with the highlights and prepared model is studied. Later this prepared information display is utilized to acknowledge the client confront picture and picture acknowledgment is then performed.

V. CONCLUSION

The face acknowledgment is a subject of machine learning. That is generally utilized for different diverse applications which includes verification and secure access control because of their uniqueness. The proposed work is committed to outline and actualize a face acknowledgment which demonstrates and utilizes the fractional or finish confront pictures with the end goal to perceive the face class. In this setting, the three-stage process is proposed to work where in the first stage the face pictures are divided into various face parts, this method is named here as the pre-preparing of pictures. After this the pictures are prepared for highlight extraction, accordingly the LDA calculation is proposed to execute. At last the neural system is proposed to perform preparing on extricated confront highlights and classes and the prepared model is utilized for perceiving the appearances. In the imminent future the proposed model is executed and their execution is given.

REFERENCES

[1] De Carrera, Proyecto Fin, and Ion Marques, "Face recognition algorithms", Master's thesis in Computer Science, Universidad Euskal Herriko, 2010.

[2] Abhishek Singh and Saurabh Kumar, “Face Recognition Using PCA and Eigen Face Approach”, B.Tech Dissertation Report, Computer Science, NIT Rourkela, May 14, 2012.

[3] Shermina J, “Illumination Invariant Face Recognition Using Discrete Cosine Transform and Principal Component Analysis”, International conference on Emerging Trends in Electrical and Computer Technology (ICETECT), 2011.

[4] M. Saraswathi and Dr. S. Sivakumari, “Evaluation of PCA and LDA techniques for Face recognition using ORL face database”, (IJCSIT) International Journal of Computer Science and Information Technologies, vol 6 (1), 2015, pp. 810-813.

[5] C.-J. Tu, L. -Y and C.-H. Yang, “Feature Selection using PSO-SVM,” International Journal of Computer Science (IAENG), vol 33, Number 1, IJCS_33_1_18.

[6] Rein-Lien Hsu, “Face Detection and Modeling for Recognition,” PhD thesis, Department of Computer Science & Engineering, Michigan State University, USA, 2002.

[7] Sanjeev Kumar and Harpreet Kaur, “Face Recognition Techniques: Classification and Comparisons”, International Journal of Information Technology and Knowledge Management July-December 2012, vol 5, No. 2, pp. 361-363.

[8] Guerreiro, Pedro Miguel Correia, “Linear Discriminant Analysis Algorithms”, Unpublished master’s thesis), Technical University of Lisbon, Portugal (2008).

[9] S. Balakrishnama and A. Ganapathiraju Linear Discriminant Analysis – A Brief Tutorial, Institute for Signal and Information Processing.

[10] Wael Abd Almageed and Yue Wu, "Face recognition using deep multi-pose representations”, In Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, pp. 1-9. 2016.

[11] Ding, Changxing, Chang Xu, and Dacheng Tao. "Multi-task pose-invariant face recognition", IEEE Transactions on Image Processing 24, Number 3, pp. 980-993, 2015.

[12] Chen, Xue-wen, Melih Aslan, Kunlei Zhang, and Thomas Huang, "Learning multi-channel deep feature representations for face recognition", In Feature Extraction: Modern Questions and Challenges, pp. 60-71. 2015.

[13] Guosheng Hu and Xiaojiang Peng, "Frankenstein: Learning deep face representations using small data", IEEE Transactions on Image Processing (2017).

[14] Hachim El Khiyari, Harry Wechsler,” Face Recognition across Time Lapse Using Convolutional Neural Networks” Journal of Information Security vol.7 No.3(2016).

ISBN 978-81-939386-1-4

Improved disease diagnosis in Healthcare using FDCT and RDWT based Medical Fusion technique

Lakshmi Devasena C
Operations and IT Department,
ICFAI Business School (IBS),
Hyderabad,
The ICFAI Foundation for Higher Education (IFHE)
(Deemed to be university u/s 3 of the UGC Act 1956)
Hyderabad - India.
devaradhe2007@gmail.com

***Abstract*—Healthcare Industry utilizes the capability of storing, analyzing and interpreting medical images to diagnose the deceases well in advance. Medical imaging is fast growing area of research. Medical images from varied modalities often provide consistent additional information for diagnosis. Several diagnostic cases require unification of complementary information from varied modalities for better study and analysis. Fused version of medical images from different modalities can deliver a single combined image that is unswerving for improved health analysis and diagnosis. Medical Image fusion is used to acquaintance relevant evidence from two or more medical images, where the resultant image will be more expounding than the input images. This research work recommends an algorithm for Medical Image Fusion of medical images from two different modalities (Magnetic Resonance Imaging (MRI) Image and Computerized Tomographic (CT)) images using Fast Discrete Curvelet Transform (FDCT) and Redundant Discrete Wavelet Transform (RDWT). This proposed technique consort the effect of both the transforms and the high quality fused image is made as output.**

Keywords— Fast discrete Curvelet Transform (FDCT); Redundant Discrete Wavelet Transform (RDWT); Medical Image Fusion

I. INTRODUCTION

In these modern days, digital data are easy to obtain and store, efficiently processed and judiciously interpreted. This bounces the new disciplines like image processing, image analysis, and computer vision. In image processing, image intensity values are used to produce an image or images in a certain sense upgraded. Image analysis deals mainly with intensity values, which are consistently enriched with additional information that helps to build a symbolic description of the content of the image. Computer vision is naturally concerned with the 2-Dimensional representation of a 3-D world; it performs conceptual reasoning on images followed by decision-making and action. These disciplines have exposed practical applications in numerous unfolding fields.

Medical Image Processing or Medical imagery field make use of images since the historical days. Many devices based on X-rays, magnetic resonance, ultrasound, scanners, etc. yield images that are subject to various processing tasks, such as quality improvement, feature enhancement and extraction, or integration of different pieces of information (image fusion). Medical images from different modalities often provide balanced information. Image fusion has a wide obligation in medical image diagnosis where medical images have a variety of modalities such as Magnetic Resonance Imaging (MRI), Computerized Tomography (CT), PET, ECT, SPECT, etc which have their own application ranges and limitations. Abundant diagnostic cases require integration of complementary information for better analysis. Fusion of different modality of medical images can provide a single composite image that is dependable for improved analysis and diagnosis.

It is practically not viable to capture all the details from single imaging modality that would guarantee clinical accuracy and robustness of the analysis. The apparent approach is to view images from multiple modalities to make a more reliable and precise assessment. This habitually requires expert readers and is often targeted at assessing details that balance the individual modalities. Some of the key modalities in clinical practice includes:

(1) Quantitative Coronary Angiography (QCA) and Quantitative Vascular Angiography (QVA),

(2) Computer Tomography,

(3) Magnetic resonance imaging such as for Angiography (MRA), Cardiac, Body and Neuro, Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI),

(4) Dual-energy X-ray absorptiometry (DXA) such as for Bone Mineral Density (BMD) and Hip Structural Analysis (HSA),

(5) Nuclear Medicine such as using Multi-Gated Acquisition Scan (MUGA) and Single Photon Emission Computed Tomography (SPECT),

(6) Ultrasound such as for Abdominal/small parts Ultrasound, Intima-Media Thickness (IMT), Echocardiography, Intravascular Ultrasound (IVUS), Duplex, Duplex Doppler, FMD (flow mediated dilation), CEUS (Contrast Enhanced Ultrasound), B-Mode and M-Mode,

(7) PET, and

(8) X-Ray imaging such as for mammography. These imaging modalities find a range of application in diagnosis and assessments of medical conditions affecting breast, brain, liver, lungs, bone marrow, mouth, stomach, teeth, intestines, bones and soft tissues.

Image fusion is a technique used to syndicate pertinent information from two or more images where the resultant image is more explanatory than the input images [1].

II. Literature Review

Significant research issue in medical image processing, particularly in information reckoning, is a fusion of multimodal information [7 to 24]. The different type of wavelet transforms used for Image Fusion are explained in [2, 3 and 4]. The Fast Discrete Curvelet transform is introduced in [5]. Existing algorithms generally use Discrete Wavelet Transform (DWT) [3] for multimodal medical image fusion [7 to 14] because DWT preserves different frequency information in a stable form and allows good localization both in the time and spatial frequency domain. Comparisons of Image Fusion techniques used so far are discussed in [15 to 18]. Recent survey on State of the art techniques in Medical Image Fusion are elaborated in [19]. Image fusion technique using Curvelet Transforms are explained in [20 and 21]. Combined Wavelet and Curvelet algorithms used for image fusion are explained in [22, 23 and 24]. A novel multimodal Medical Image Fusion (MIF) method based on improved Contourlet Transform (CNT) for multi-sensor, spatially registered, multi- resolution medical images is projected in [25]. An application of Neural Network (NN), a novel feature-level multi-focus image fusion technique has been described and implemented in [26], which fuses multi-focus image using classification. In the pixel level, fusion between the medical images using a Stationary wavelet transform algorithm is presented in [27]. Pixel-Level Image Fusion for satellite images using Kalman Algorithm is presented in [28]. Different types of image fusion algorithms based on the discrete cosine transform (DCT) were developed and their performances were evaluated in [29]. Different modality images are fused using the maximum fusion rules based on the Multi-Wavelet and wavelet transforms in [30]. An improved block based feature level image fusion technique using lifting wavelet transform and neural network (BFLN) is proposed in [31]. A novel multimodal Medical Image Fusion (MIF) method, based on Ripplet Transform Type-I (RT) for multi-sensor, spatially registered, multi-resolution medical images is recommended in [32]. Detection of Throat Cancer Types is experimented using a Fusion Technique Based on Image - Statistical Analysis in [33]. Cohesive PCA & DCT Based Fusion using Reliability Verification & Non-Linear Enhancement is done in [34]. Convolution of Meridian Distributions is used for Medical Image Fusion in [35].

Most of these existing works are based on DWT. However, one of the major drawbacks of DWT is that the DWT transformation does not provide shift invariance. This causes a key change in the wavelet coefficients of the image even for negligible shifts in the input image. In medical imaging, it is important to preserve the exact location of the wavelet information; but shift variance may lead to inaccuracies. For example, in medical image fusion edge information should be preserved, but DWT based fusion techniques may produce peculiarities along the edges, which may lead to inaccurate fused image. To avoid this shift invariance, the proposed algorithm uses Redundant Discrete Wavelet Transform [6] combined with Fast Discrete Curvelet Transform, to acquire better quality and accurate fused image of MRI & CT images.

III. Methodology

A. *Fast Discrete Curvelet Transform*

The Curvelet transform is formed from Wavelet transform by generalizing its higher dimensional to represent images at different angles and different scales. Curvelets have the benefit of two unique mathematical properties, namely: 1) Curved singularities can be approximated with very few coefficients and in a non-adaptive manner. 2) Curvelets remain articulate waveforms under the action of the wave equation in a smooth medium.

In curvelet transform, the complete image is segmented into small overlapping tiles and subsequently ridgelet transform is applied on each tile [5].

Algorithm for first generation curvelet is given below:

Step1. Decompose the input image into 3 sub bands using Sub band filtering by applying additive wavelet transform.

Step2. Accomplish tiling on each of the sub bands.

Step3. Implement Discrete Ridgelet Transform on each of tile on all the sub bands.

Tiling involves dividing the sub bands 1 and 2 of the transformed image into overlapping tiles, resulting in smaller dimensions to transform the curved lines to straight lines thus avoiding the edge effects. Segmentation is performed to approximate curved lines into overlapping tiles. Discrete Ridgelet transform is performed on the segmented tiles. This is done by applying 1-Dimensional discrete wavelet transform to the slices of the radon transform. Ridgelet transform helps in shape detection. The first generation curvelet transform is more complex involves a series of steps [5].

Due to its complexity, the second generation curvelet is much desired. Thus Second generation does not use the ridgelet transform, and hence reduces the redundancy.

1. Wrapping Algorithm

i. Execute FFT on the original image.

ii. Divide FFT into a collection of tiles

iii. For each tile apply

a. Convert tile to the origin.

b. Cover parallelogram shaped support of tile around the rectangle.

c. Take inverse FFT of wrapped one

d. Add it to collection of curvelet coefficients already acquired.

2. Inverse Wrapping Algorithm

i. For each curvelet coefficient array,

a. Take FFT of the array.

b. Unwrap rectangular support to original orientation shape.

c. Decode it back to the original position

d. Store the translated array

ii. Add all the translated curvelet arrays

iii. Take inverse FFT to restructure the image.)

B. Redundant Discrete Wavelet Transform

Redundant discrete wavelet transform (RDWT) [2, 3], another variant of wavelet transform, is used to overcome the shift variance problem of DWT. It has been applied in diverse signal processing applications, but it is not well explored in the field of medical image fusion. RDWT can be considered as an approximation to DWT that removes the down-sampling operation from traditional disparagingly sampled DWT and produces an over-complete representation and then provides noise free sub-band relationship. Since the spatial sampling rate of RDWT is fixed across scale, it is shift invariant. Alike DWT, RDWT and Inverse RDWT (IRDWT) of two dimensional image or three dimensional volume data can be attained by computing each dimension individually and it will be the same size as the input image/data [6].

C. Proposed FDCT and RDWT based algorithm

Let I1 and I2 be the MRI and CT brain images. First FDCT is applied and the images are reconstructed as RI1 and RI2. Then three levels of RDWT decomposition are applied on both the reconstructed images to obtain the detail and approximation wavelet bands. Let Ia1, Iv1, Id1, and Ih1 be the RDWT sub-bands from RI1 and Ia2, Iv2, Id2, and Ih2 be the corresponding RDWT sub-bands from RI2 image.

To conserve the features from both the images, coefficients from the approximation band of RI1 and RI2 are averaged, as given in Equation 1.

$$I_a^F = \text{mean}\,(I_a^1, I_a^2) \quad (1)$$

where, IaF is the approximation band of the fused image. For the three detailed sub-bands, each band is divided into blocks of size 3 x 3 and the entropy of each block is calculated using Equation 2.

$$e_i^{jk} = ln\sqrt{\left(\frac{\mu_i^{jk} - \sum_{x,y=1}^{3,3} I_i^{jk}(x,y)}{\sigma_i^{jk}}\right)^2 / m^2} \quad (2)$$

where ‘j’ (=a, v, d, h) represents the sub-bands, ‘m’ is block size (here it is 3), ‘k’ is block number, and ‘i’ is used to differentiate the two multimodal images (RI1 of one particular modality) and RI2 (of another modality). □ijk and μijk are the standard deviation (SD) and mean of the RDWT coefficients of the jth sub-band of kth block from ith image. Using the entropy values, the detail sub-bands for the fused image IvF, IdF, and IhF are generated using Equation 3. For fused image block IFjk, RDWT coefficients from RI1 is selected if the entropy of block from RI1 image is greater than the corresponding block from the RI2 image, otherwise I2jk is selected.

$$I_F^{jk} = \begin{cases} I_1^{jk}, & if \ (e_1^{jk}) > (e_2^{jk}) \\ I_2^{jk}, & otherwise \end{cases} \quad (3)$$

Then, Inverse RDWT transform is applied to the fused sub-bands to produce the fused medical image IF.

$$I^F = \text{IRDWT}\,(I_a^F, I_v^F, I_d^F, I_h^F) \quad (4)$$

Finally inverse RDWT is applied to generate the final region based fused image. The overall process of the proposed technique is shown in Fig. 1.

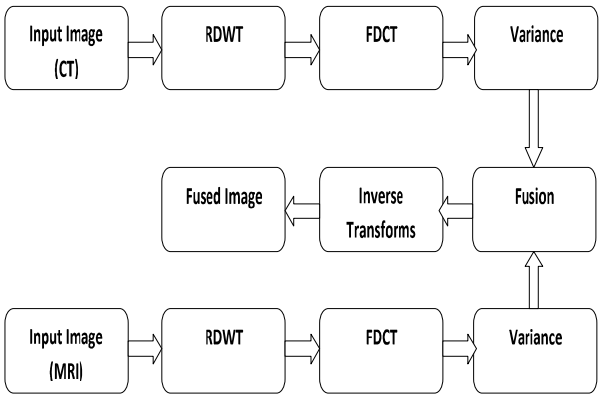
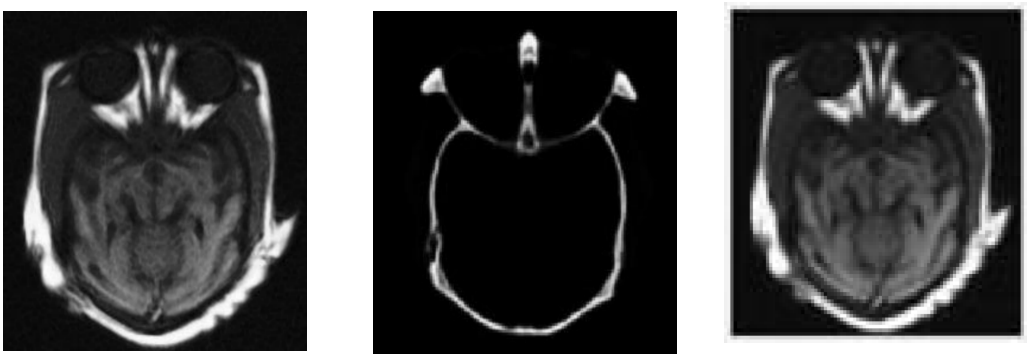


Fig. 1. Overall Processof proposed algorithm

IV. Results and Discussion

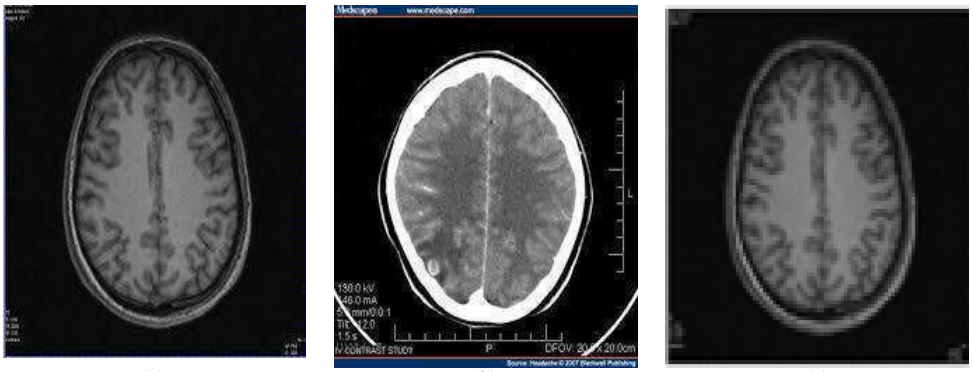
The proposed algorithm has been investigated using Matlab and experimented with MRI and CT scan images of different part of the body. The proposed technique works as follows: First the MRI image is given as first image and CT scan image is given as a second input image. RDWT is applied to the input images to get the reconstructed image. In a similar way, second generation Fast Discrete Curvelet Transform is applied to the input images. Then the variances are fused. After which, for each curvelet array, Inverse Fast Discrete Curvelet Transform is applied. Finally, Inverse RDWT is applied to generate the final region based fused image.

The experimented results are shown in Fig. 2 to Fig. 5. In Fig. 2, MRI and CT scan brain images are taken as input and the proposed combined FDCT and RDWT resultant fused image is shown in Fig. 2 (c). In fig. 3, another MRI and CT scan Brain images are taken as input and the resultant fused image is shown in Fig. 3 (c). In Fig. 4, MRI and CT scan Spinal card images are taken as input and the combined FDCT and RDWT fused image using proposed technique is shown in Fig. 4 (c). Finally, in Fig. 5, MRI and CT scan HIP images are taken as input images and the Region based fused image of proposed combined FDCT and RDWT is shown in Fig. 5 (c).



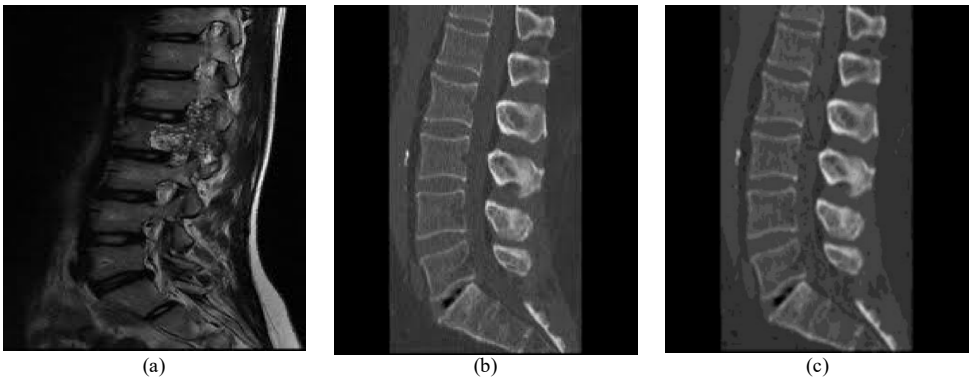
(a) (b) (c)

Fig. 2. a) MRI Brain Image b) CT Scan Brain Image c) Region based Fused Brain Image of combined FDCTand RDWT.



(a) (b) (c)

Fig. 3. a) MRI Brain Image b) CT Scan Brain Image c) Region based Fused Brain Image using combined FDCTand RDWT.



(a) (b) (c)

Fig. 4. a) MRI Spinal Card Image b) CT Scan Spinal Card Image c) Region based Fused Spinal Card Image of combined FDCTand RDWT.



(a) (b) (c)

Fig. 5. a) MRI Hip Image b) CT Scan Hip Image c) Region based Fused Hip Image of combined FDCT and RDWT.

V. CONCLUSION

Medical images from diverse modalities habitually provide balanced information. Several diagnostic cases necessitate integration of balanced information for better analysis. Medical image fusion of different modalities can provide a single amalgamated image that is trustworthy for enhanced analysis and diagnosis. In this research work, a

new algorithm for Medical Image Fusion of MRI and CT Medical images using Redundant Discrete Wavelet Transform (RDWT) and Fast Discrete Curvelet Transform (FDCT) is being proposed and experimented with MRI and CT Scan images of different part of the body like, brain, spinal card, and hip images.

ACKNOWLEDGMENT

The author express her heart felt gratitude to the management of The ICFAI Foundation for Higher Education (IFHE) for the support and motivation extended.

REFERENCES

[1] Simone G, Farina A, Morabito, Serpico S, and Bruzzon L, "Image fusion techniques for remote sensing applications," Inf fusion. vol 3, pp. 3-15, 2002.

[2] I. Daubechies, "Ten lectures on wavelets," Society for Industrial and Applied Mathematics. 1992.

[3] J. Fowler, "The redundant discrete wavelet transform and additive noise," IEEE Signal Processing Letters. vol 12(9), pp. 629–632, 2005.

[4] G. Strang and T. Nguyen, "Wavelet and filter banks," Wellesly-Cambridge Press. 1996.

[5] Emmanuel Cand`es, Laurent Demanet, David Donoho and Lexing Ying, "Fast Discrete Curvelet Transforms," Multiscale Model. Simul. Society for Industrial and Applied Mathematics, vol.5(3), pp. 861–899, 2006.

[6] Richa Singh, Mayank Vatsa and Afzel Noore, "Multimodal Medical Image Fusion using Redundant DiscreteWavelet Transform," Proceedings: In International Conference on Advances in Pattern Recognition, 2009.

[7] Q. Guihong, Z. Dali, and Y. Pingfan, "Medical image fusion by wavelet transform modulus maxima," Optics Express. vol. 9(4), pp. 184 – 190, 2001.

[8] Y. Kirankumar and S. Devi, "Transform-based medical image fusion," International Journal of Biomedical Engineering and Technology. vol. 1(1), pp. 101 – 110, 2007.

[9] Y. Licai, L. Xin, and Y. Yucui, "Medical image fusion based on wavelet packet transform and self-adaptive operator," In Proceedings of International Conference on Bioinformatics and Biomedical Engineering. pp. 2647 – 2650, 2008.

[10] D. Townsend and T. Beyer, "A combined PET/CT scanner: the path to true image fusion," The British Journal of Radiology. vol. 75, pp. 24 – 30, 2002.

[11] Z. Wang and Y. Ma, "Medical image fusion using m-PCN. Information Fusion," vol. 9(2), pp. 176–185, 2008.

[12] H. Zhang, L. Liu, and N. Lin, "A novel wavelet medical image fusion method," In Proceedings of International Conference on Multimedia and Ubiquitous Engineering, pp. 548–553, 2007.

[13] R.J.Sapkal, S.M.Kulkarni, "Image Fusion based on Wavelet Transform for Medical Application," International Journal of Engineering Research and Applications. vol. 2(5), pp.624-627, September- October 2012.

[14] Deepali Sale, Madhuri Joshi, Varsha Patil, Pallavi Sonare, and Chaya Jadhav, "Image Fusion For Medical Image Retrieval," International Journal of Computational Engineering Research. Vol. 3(8), pp. 1 – 9, August 2013.

[15] Chu-Hui Lee and Zheng-Wei Zhou, "Comparison of Image Fusion based on DCT-STD and DWT-STD," Proceedings of International Multi-Conference of Engineers and Computer Scientists, vol. 1, March 2012, Hong Kong.

[16] Sandeep Kumar, Yash Kumar Sharma and Mahua Bhattacharya, "Curvelet Based Multi-Focus Medical Image Fusion Technique: Comparative Study with Wavelet Based Approach".

[17] Sweta Mehta, Bijith Marakarkandy and Jaydev Shilu, "Comparison of Fusion Algorithms for Fusion of CT and MRI Images," International Journal of Computer Applications, 2013.

[18] Y. Kiran Kumar, "Comparison of Fusion Techniques applied to Preclinical Images: Fast Discrete Curvelet Transform using Wrapping Technique & Wavelet Transform," Journal of Theoretical and Applied Information Technology. pp. 668 – 673, 2005 – 2009.

[19] A.P. James, B. V. Dasarathy, "Medical Image Fusion: A survey of the state of the art," Information Fusion. 2014.

[20] Vetrivelan, P and Kandaswamy, A, "Fusion of Medical Images using Curvelet Transform," International Journal of Recent Scientific Research. Vol. 3(7), pp. 621 - 626, July 2012.

[21] Sweta Mehta and Bijith Marakarkandy, "CT and MRI Image Fusion using Curvelet Transform," Journal of Information, Knowledge and Research in Electronics and Communication Engineering. vol. 2(2), pp. 848 – 852, November 2012 – Ocober 2013.

[22] Roshna J Sapkal and Sunita M Kulkarni, "Innovative Multilevel Image Fusion Algorithm using Combination of Transform Domain and Spatial Domain Methods with Comparative Analysis of Wavelet and Curvelet Transform," International Journal of Computer Applications. vol. 66(21), pp. 41 – 47, March 2013.

[23] T. Sumathi and M. Hemalatha, "An Innovative Image Fusion Algorithm based on Wavelet Transform and Discrete Fast Curvelet Transform," Central European Journal of Computer Science.vol.1(3), pp. 329-340, September 2011.

[24] Olivier Riou and Pierre Duhame, "Fast Algorithms for Discrete and Continuous Wavelet Transforms," IEEE Transactions on Information Theory, vol. 38(2), pp. 569 – 586, March 1992.

[25] Rishamjot kaur and Harmandeep Singh, "Fusion of Multimodality Medical Images Based on Multi-Scale Decomposition using Contourlet Transform," International Journal of Research in Engineering & Applied Sciences, vol. 2(7), pp. 11 – 20, July 2012.

[26] G. Mamatha, Shaik Abdul Rahim and Cyril Prasanna Raj, "Feature-Level Multi-Focus Image Fusion Using Neural Network and Image Enhancement," Global Journal of Computer Science and Technology Graphics & Vision, vol. 12(10), June 2012.

[27] Sanket Choudhary and Yash Kshirsagar, "Medical Image Fusion Using Stationary Wavelet Transform," Int. J. of Emerging Trends in Electronics and Computer Science, vol. 2(1), pp. 1 – 5, January 2013.

[28] S. A. Quadri and Othman Sidek, "Pixel-Level Image Fusion using Kalman Algorithm," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 6(2), pp. 75 – 86, April, 2013.

[29] VPS Naidu, "Discrete Cosine Transform based Image Fusion Techniques," Journal of Communication, Navigation and Signal Processing, Vol. 1(1), pp. 35-45, January 2012.

[30] Kusum Rani, and Reecha Sharma, "Study of Image Fusion using Discrete wavelet and Multiwavelet Transform," International Journal of Innovative Research in Computer and Communication Engineering vol. 1(4), pp. 795 – 799, June 2013.

[31] Chetty. Manna Sheela Rani and V.VijayaKumar, "Block based Image fusion technique using Lifting wavelet transform and Neural networks on Medical Images," International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 2(5), pp. 980 – 986, October 2012.

[32] S. Das, M. Chowdhury, and M. K. Kundu, "Medical Image Fusion Based on Ripplet Transform Type-I," Electromagnetics Research B, Vol. 30, 355 – 370, 2011.

[33] Adnan Al-Bashir and Bassam Al-Naami, A Fusion Technique Based on Image - Statistical Analysis for Detection of Throat Cancer Types, Jordan Journal of Mechanical and Industrial Engineering, Volume 4, Number 6, December 2010, Pages 677 – 684.

[34] Shaveta Mahajan and Arpinder Singh, Integrated PCA & DCT Based Fusion Using Consistency Verification & Non-Linear Enhancement, International Journal Of Engineering And Computer Science,Volume 3 Issue 3 March, 2014 Page No. 4030-4039.

[35] Mayank Agrawal, Panagiotis Tsakalides, and Alin Achim, Medical Image Fusion using the Convolution of Meridian Distributions, 32nd Annual International IEEE EMBS Conference, 2010.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A New Text Mining Approach for Web Content Search

Thayyaba Khatoon Mohammed[1]
Malla Reddy College of Engineering and Technology
Hyderabad, Telangana State,
Indiathayyaba.khatoon16@gmail.com

Chandrasekar V[2]
Malla Reddy College of Engineering and Technology
Hyderabad, Telangana State, India
drchandru86@gmail.com*

A Govardhan[3]
Rector, JNTUH
Kukatpally, Hyderabad-500085, India
govardhan_cse@jntuh.ac.in

Abstract*— Text-mining (TM) refers generally to the practice of extracting attractive and non-trivial information and facts from unstructured text. TM includes several computer science (CS) regulations with a strong direction towards artificial intelligence (AI) in general, including but not limited to pattern recognition (PR), neural networks (NN), natural language processing (NLP), information retrieval (IR) and machine learning (ML). A significant variation with search is that search requires a user to identify what he or she is looking for while TM attempts to realize information in a model that is not known earlier.TM is mainly motivating in domains where users have to invent new information. This is the case, for e.g., in criminal enquiries and legal findings. Such examinations require 100% evoke, i.e., users can not meet the expense of missing relevant data. In distinction, a user searching the internet for background information using a benchmark search engine (SE) simply requires any data as long as it is reliable. A different technique is to combine standard significance ranking with adaptive filtering (AF) and interactive visualization (IV) that is based on characteristics that have been mined earlier.

Keywords*— Artificial intelligence, Neural networks, Machine learning, Interactive Visualization, Search Engine, Adaptive Filtering, Text Mining.

I. INTRODUCTION

Within the area of expertise of TM, now and then called text analytics, several attractive technologies such as computers, IT, PR, statistics, advanced mathematical techniques, AI, visualization, and IR. The information bang of modern times will persist at the same rate **[1, 2]**. TM techniques play a crucial role in the upcoming years in this lifelong process. Due to continuing globalization there is also much interest in multi-language text mining (MLTM) TM: the attaining of insights in ML collections. MLTM is much more difficult that it appears as in addition to differences in words and character sets, TM makes concentrated use of data as well as the linguistic properties of a language. There are many essential hypotheses about capitalization and tokenization that would not work for other languages. When TM methods are used on non-English data collections supplementary challenges have to be concentrated **[4, 5]**.

TM is about investigating unstructured information and extracting relevant patterns or models and uniqueness. Using these models and characteristics better search results and deeper data analysis is possible; giving quick IR otherwise it would remain hidden. The field of data mining (DM) is better known than that of TM. A good e.g., of DM is the analyzing of operational details contained in relational databases, such as debit card transactions or credit card payments. To such operational diverse supplementary information can be provided: date, location, age of card holder, salary, etc. With support of this information patterns of behavior can be determined However, 95% of all information is unstructured information, and both the proportion and the total amount of unstructured information raise daily.

Only a small amount of information is stored in a structured format in a relational database. The greater part of information that users work on every day is in the form of text documents (TD), E-Mails/ multimedia files (speech, video and photos). Searching analysis using database (DB) or DM methods of this information is not possible, as these procedures work only on structured information (SI). It is easier to manage, share, search, organize, and to generate reports so on, for computers as well as users, therefore the wish is to give structure to unstructured information. This allows computers and public to better manage the data, and allow known procedures and methods to be used.

TM, using manual procedures, was use first during the 1980s. It hastily became obvious that these manual procedures were labor demanding and therefore costly **[6]**.It also cost too much time to manually process the already-growing amount of information. Over time there was growing success in creating applications to mechanically process the data, and in the previous ten years much development has seen.

Currently the study of TM worries the growth of various mathematical, statistical, and methods which allow mechanical analysis of unstructured information as well as the extraction of high quality and appropriate data, and to make the text as a whole better searchable. High excellence refers here, irrelevance and the obtaining of new and interesting approaches. A TD contains characters that jointly form words, which can be pooled to structure phrases. They are all syntactic assets or properties that collectively symbolize defined categories, concepts, meanings. TM must distinguish, extract and use all this information. Using TM, instead of searching for words, we can search for linguistic (scientific study of language) word patterns and this is therefore searching at a higher level.

II. SEARCHING UNSTRUCTURED INFORMTATION

What turns out exactly when somebody uses a computer program to search unstructured text? Computers are digital apparatus with limited capabilities. Computers manage best with figures, especially complete (whole) numbers, also known as integers, if it has to be really fast. Public are analogue, and individual language is also analogue, full of irregularity, obstruction, errors and exceptions. If public search for something then they often think in concepts and semantics, all regions in which a machine can't straightly transact with. For machines to be able to make a computationally proficient search in a huge amount of text, the difficulty needs first to be transformed to a numerical problem that a computer can associate with. This directs to very large storages containing many numbers in which numbers characterizing search terms are compared with numbers characterizing information and documents. This is the basic standard that our field concerns itself with: translation of information that users can work with into information that a computer can work with, and then convert the result back into a form that public can understand.

This expertise exists ever since 1960's. One of the initial experts working in this field was Gerard Salton who was an expert in earlier days; both jointly with other experts complete one of the initial texts SE. Each incident of a word in the text was entered in a keyword index. Searching was then done in the index, analogous to the index at the back of a book but with many more words and much earlier.. With procedures such as B-trees and hashing, it was probable to speedily and proficiently make a list from all documents containing a word or a Boolean combination of words **[7]**.

Documents and search terms were transformed to vectors and contrasted using the cosine distance among them: how lesser the cosine distance, how additional the search term and the document communicated. This was an efficient technique to decide the relevance of a document from the search expression.This was called the vector space model, and is still used today by some applications. Later on, a variety of other techniques used for searching and particular, to the combination of the relevance.

There are many search procedures with good-sounding names such as: directed and non-directed proximity, fuzzy, wildcards, semantically, taxonomies, conceptual, etc. Examples of normally known relevance defining procedures are: term-based frequency ranking (TBFR), the page-rank algorithm (PRA), and probabilistic ranking (PR). Because these days there is so much information digitally available and because it is now often crucial to directly react on present activities, new procedures are necessary to keep up with the continuously rising amount of unstructured information. In addition, people will have different causes for searching large amount of data and different goals to find, and those distinctions require an alternative advances.

III. SEARCHING AND FINDING

Fraud researchers or public prosecutors don't only want the best documents; they want all probable appropriate documents. In addition, in an internet SE everyone does their best to get to the top of the results list; search engine optimization has in itself become a skill. This is done by using synonyms and code names, and relatively often these are common words that are used so often that a search cannot be done without returning millions of knocks (hits). TM can offer a solution to find the appropriate information.

Fraud researchers also have another frequent difficulty: at the beginning of the analysis they do not know accurately what they must search for. They do not know the code names or synonyms, or they do not exactly know which companies, account numbers, persons, or amounts must be searched for. Using TM it is likely to recognize all these types of properties from their linguistic role, and then to categorize them in a structured manner to present them to the user. It then becomes very simple to investigate the found companies further.

Sometimes the difficulties facing by a researcher go a little deeper: they are searching without really knowing what they are searching for. TM can be used to discover the words and subjects important for the examination; the computer searches for specified patterns in the text: "who paid to whom", "who talked to whom", etc. These types of patterns can be predictable using language technology and TM, and extracted from the text and presented to the researcher, who can then quickly decide the legal transactions from the suspect ones.

An example: If the HDFC bank transfers money to the AXIS then that is a normal transaction. But if "Norman" transfers money to Kumaran Enterprises Inc. then that may be doubtful. TM can recognize these natures of patterns, and further searches can be made on the words in those patterns using normal search procedures to further recognize and examine details. The obtaining of new insights is also called fortune. TM can be custom-made very efficiently to obtain new but often vital insights necessary to progress in an examination.

Therefore, the TM helps in the search for data by using models for which the values of the elements are not exactly known earlier. This is analogous with mathematical functions in which the variables and the statistical distribution of the variables are not always known. Here the heart of the difficulty can be seen as a conversion problem from human language to mathematics. The better the mathematical conversion, the better excellence of the TM will be.

IV. INFORMATION VISUALIZATION

TM is often stated as information visualisation (IV). This is because visualisation is one of the practical possibilities after unstructured information has been structured. To be able to make these sorts of visualizations the features must be structured or planned, and that is accurately the area in which TM expertise can help: by structuring unstructured information it is possible to visualise the data and more rapidly obtain new insights **[8]**

Washington, The Pacific Northwest, July 16th, 2001 - ZyLAB, the developer of document imaging and full-text retrieval software, has donated a PACS system to the government of U.S **[3]**. "We have been working closely with the World Health Organization, US International for the last 4

years now," said, CEO of ZyLAB Technologies. “The USA government faced hard job with the ZyLAB system will be of tremendous assistance to them. Unfortunately, the U.S has scarce resources to procure advanced imaging and archiving systems to help them in this job, full operational PAC system was donated to them."

An example is the following text:

In that text, the following entities and attributes can be found in the Table – 1:

Table – 1: Entities and Attributes of Above Text

Places	Washington
Countries	The Pacific Northwest, U.S
Persons	Gerald Gahima, David
Function titles	CEO
Data	July 16th, 2001
Organisations	Government of U.S, WHO, Criminal Justice Source Center, American-Canadian (AC) helper group
Companies	ZyLAB, ZyLAB Technologies BV
Products	PACS

Let’s assume that there are various documents containing this type of automatically found structured properties; then the documents could not only be presented in tabular form, but also for e.g., in a tree structure in which the document could be structured on occurrences per land and then on occurrences per administration. The principle can also be used to dynamically imagine a tree structure, which would then appear as shown in the Fig – 1.

With the assist from these kinds of visualization methods it is possible to gain a quicker and better insight into difficult data collections, particularly if it involves large groups of unstructured information that can be mechanically structured using DM.

With the assist from these kinds of visualization methods it is possible to gain a quicker and better insight into difficult data collections, particularly if it involves large groups of unstructured information that can be mechanically structured using DM

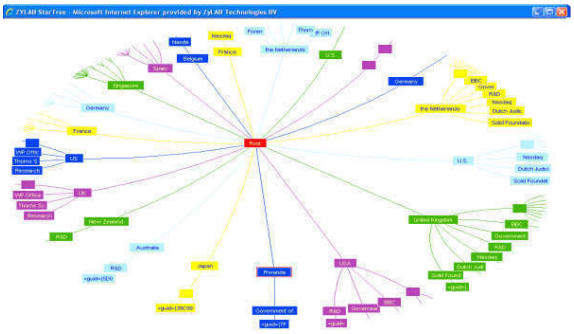


Fig 1: Hyperbolic Tree visualisation of a tree structure

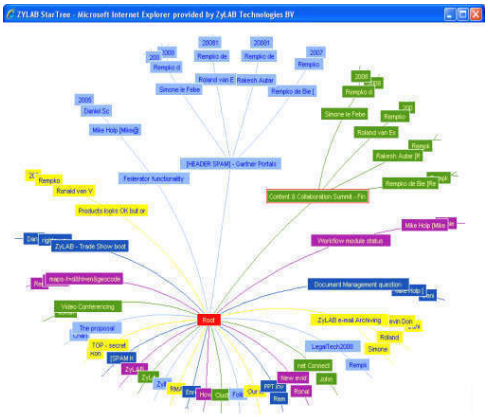


Fig 2: E-mail visualisation using a Hyperbolic Tree

.

Advantages of Structured and Analysed Data

In addition to the visualization, various other search extensions are possible when the data has been structured and has Meta details.

Here is a concise list:

Details are easier to organize in folders. It is easier to filter data on specified Meta details, when searching Details can be contrasted, and linked using the Meta details. It is probable to sort, group and prioritize the documents using any of the attributes. Particulars can be grouped using the Meta elements.With the assist of Meta details duplicates and almost-duplicates can be sensed.

These can then be deleted. Categorizations can be copied from the Meta elements. It is feasible to search the Meta details from already found documents. Statistical reports can be made on the basis of the Meta details.

V. Conclusion

Although changes in the authorized world are always evolutions, there is definitely a probable function for TM in Electronic-Discovery and Electronic-Disclosure. Data sets are just reaching large and must be re-examined serially. Groups need to be pre-examined and pre-documented. Re-

examinations can be implemented more efficiently and goals can be made easier. The challenge will be to encourage courts of the rightness of these new tools. Therefore, a hybrid advance is recommended where computers make the initial selection and classification of documents and research directions and human reviewers and researchers implement quality control and valuate the survey suggestions. By doing so, computers can focus on recall and human being can focus on accuracy. There are many other applications where this approach has led to both more competence but also to acceptance of the technology by the public.

References

[1] Allan, James (Editor), (2002). *Kluwer Academic Publishers.*

[2] Andrews, Whit and Knox, Rita (2008). *Gartner Research Report, ID Number: G00161178. Gartner, Inc*

[3] Baron, Jason R. (2005). *Sedona Conference Journal. Vol. 6, 2005*

[4] Berry, M.W., Editor (2004). *Springer- Verlag.*

[5] Berry, M. W. and Castellanos, M. Editors (2006). *Springer-Verlag.*

[6] Bilisoly, Roger (2008). *Wiley Series on Methods Applications in Data Mining. Applications in Data Mining.*

[7] Blair, D.C. and Maron, M.E. (1985). *Communications of the ACM, Vol. 28, No. 3, pp. 289-299.*

[8] Card, Stuart K., Mackinlay, Jock D., and Shneiderman, Ben, Editors (1999). *Morgan Kaufmann Publishers*.

[9] http://www.mimuw.edu.pl/~son/datamining/materials/search_engine.pdf

[10] Hwee-Leng Ong, Ah-Hwee Tan, Jamie Ng, Hong Pan and Qiu-Xiang Li.(2001),"FOCI : Flexible Organizer for Competitive Intelligence", In Proceedings, Tenth International Conference on Information and Knowledge Management (CIKM'01), p p. 523-525, Atlanta, USA, 5-10

[11] XiQuan Yang, DiNa Guo, XueYa Cao and JianYuan Zhou (2008), "Research on Ontology-based Text Clustering", Third International Workshop on Semantic Media Adaptation and Personalization China,, IEEE Computer Society , 141-146.

[12] Emilio Sanchis, Davide Buscaldi, Sergio Grau, Lluis Hurtado and David Griol (2006)," SPOKEN QA BASED ON A PASSAGE RETRIEVAL ENGINE", Proceedings of IEEE international confernce, Spain, 62-65.

[13] Pak Chung Wong, Paul Whitney and Jim Thomas," Visualizing Association Rules for Text Mining", ", International Conference, Pacific Northwest National Laboratory, USA, 1-5

[14] Journal of Universal Computer Science, vol. 19, no. 3 (2013), 406-427 submitted: 22/9/11, accepted: 28/1/13, appeared: 1/2/13 © J.UCS

[15] Maziarz M., Piasecki M., Szpakowicz S.: Approaching plWordNet 2.0. In Proc. of the 6th Global Wor dnet Conference, Matsue, Japan, January 2012.

[16] Amr Ahmed and Eric P. Xing. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. In UAI, pages 20–29, 2010.

.

ISBN 978-81-939386-1-4

A Comprehensive review on Density-Based clustering algorithm in data mining

[1]R.Sujatha, [2]Dr.V.Chandrasekhar

Dept of CSE, Malla Reddy College of Engineering & Technology, Telangana

Abstract

The problem of detecting clusters of points in data is challenging when the clusters are of different size, density and shape. Many of these issues become even more significant when the data is of very high dimensionality and when it includes noise and outliers. Clusters are identified by looking at the density of points. This paper gives an overview on various density based cluster algorithms – DBSCAN,OPTICS, and DENCLUE. These algorithms are particularly suited to deal with large datasets, with noise, and are able to identify clusters with different sizes and shapes. In machine learning and data analytics clustering methods are useful tools that help us visualize and understand data better.

Keywords:Density,Clustering,Minpoints,DBSCAN,OPTICS,DENCLUE

1. Introduction

DBSCAN is a density based clustering algorithm, it is focused on finding neighbors by density (MinPts) on an 'n-dimensional sphere' with radius ε. A cluster can be defined as the maximal set of 'density connected points' in the feature space.[1] DBSCAN algorithm is that, for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points, that is, the density in the neighborhood has to exceed some predefined threshold.[2] This algorithm needs three input parameters:

- k, the neighbour list size
- Eps, the radius that delimitate the neighborhood area of a point (Epsneighbourhood)
- MinPts, the minimum number of points that must exist in the Eps-neighbourhood.

1.1 Parameter estimation

The parameter estimation is a problem for every data mining task. To choose good parameters we need to understand how they are used and have at least a basic previous knowledge about the data set that will be used.[3]

eps: if the eps value chosen is too small, a large part of the data will not be clustered. It will be considered outliers because don't satisfy the number of points to create a dense region. The eps should be chosen based on the distance of the dataset (we can use a k-distance graph to find it), but in general small eps values are preferable.
minPoints: As a general rule, a minimum minPoints can be derived from a number of dimensions (D) in the data set, as minPoints ≥ D + 1. Larger values are usually better for data sets with noise and will form more significant clusters. The minimum value for the minPoints must be 3, but the larger the data set, the larger the minPoints value that should be chosen.

To clusters a dataset, our DBSCAN implementation starts by identifying the k nearest neighbors of each point and identify the farthest k nearest neighbor (in terms of Euclidean distance) . [4] The average of all this distance is then calculated. After that, for each point of the dataset

the algorithm identifies the directly density-reachable points and classifies the points into core or border points.The clustering process is based on the classification of the points in the dataset as core points, border points and noise points, and on the use of density relations between points to form the clusters.[5]

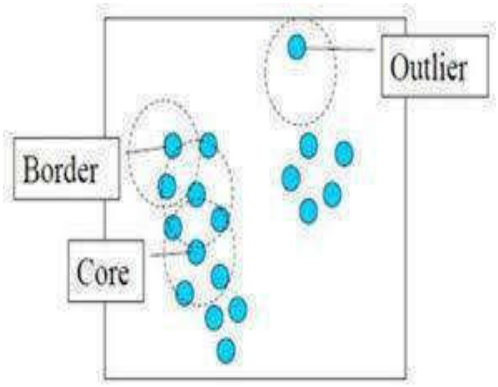


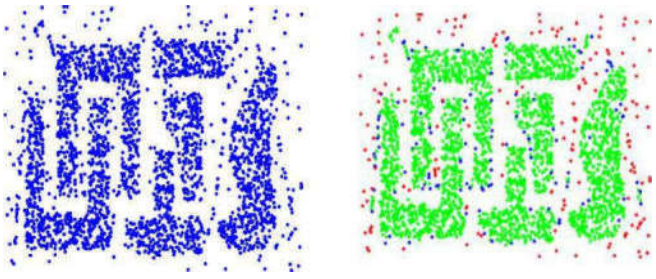
Fig.1 Diagram Representation of classification points

A point is a core point show in figure.1 if it has more than a specified number of points (MinPts) within Eps—these are points that are at the interior of a cluster.

A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point.

A noise point is any point that is neither a core point nor a border point

Example:



Original data **Point types: core, border and Outliers**

Fig.2 Representation of classification points

Density-based clustering Algorithms mainly include three techniques:

- DBSCAN which grows clusters according to a density-based connectivity analysis.
- OPTICS extends DBSCAN to produce a cluster ordering obtained from a wide range of parametersettings.
- DENCLUE Clusters objects based on a set of density distributionfunctions.[7]

2. ***Density Based Clusters*** Clusters may be looked at as dense regions in thedata space; where clusters are separated by a sparseregion containing “relatively few” data.

Given thisassumption, a cluster can either be of “regular” or“arbitrary” shape.Density based clustering is todetect clusters of non-spherical or arbitrary shapes. Some of the common density based clusteringtechniques are DBSCAN, OPTICS, VDBSCAN,DVBSCAN, DBCLASD, ST-

DBSCAN andDENCLUE. One of the earliest density based clustering methods is *Density-Based Spatial Clustering of Applications with Noise (DBSCAN).[8]*

DBSCAN discovers high density regions inspatial databases with noise and creates clusters out ofthem.It uses the concept of **density reachability** and **density connectivity**.

Density Reachability - A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbors who are within distance ε. **Density Connectivity** - A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbors and both the points "p" and "q" are within the ε distance. This is chaining process. So, if "q" is neighbor of "r", "r" is neighbor of "s", "s"

ISBN 978-81-939386-1-4

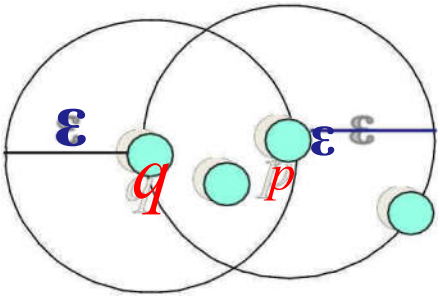
is neighbor of "t" which in turn is neighbor of "p" implies that "q" is neighbor of "p".

Density reachability

Directlydensity-reachableAn object q as shown in figure.3 is directly density-reachable fromobjectqifqiswithintheε-neighborhoodof p and p is a coreobject.[9]

*q*is directly density-reachable from*p*

*p*is not directly density-reachablefrom *q*



MinPts = 4

Fig.3 Directly density-reachable

Density-Reachable (directly andindirectly):

Density-Reachable: An object pas shown in figure.4 is density-reachable from q w.r.t ε and MinPts if there is a chain of objects $p_1,\ldots,p_n$, with $p_1=q$, $p_n=p$ such that $p_{i+1}$is directly density-reachable from p_i w.r.t ε and *MinPts* for all1<= *i* <= *n*

A point *p* is directly density-reachable

from*p*2 *p*2is directly density-reachable

from*p*1 *p*1is directly density-reachable

from*q p* *p*2 *p*1 *q* form achain

p is (indirectly)density-reachable from *q*

q is not density-reachable from *p*

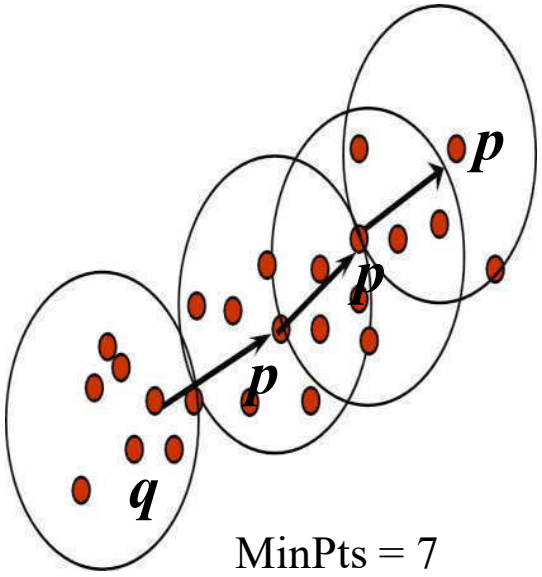


Fig.4 Representation of Density-Reachability

3. DBSCAN ALGORITHM:

- Select a point ***p.***
- Retrieve all points density-reachable from***p*** with respect to ε and MinPts.
- If ***p*** is a core point, a cluster isformed.
- If ***p*** is a border point, no points are density-reachable from ***p*** and it visits the next point of thedatabase.
- Continue the process until all the points have been processed.[10]

Advantages of DBSCAN

Clusters of arbitrary shape can be detected No prior knowledge about the number of clusters is required There is a notion of noise (objects not belonging to any cluster) Only two (2) input parameters (ε – *radius* and*MinPts*– minimum number of points) and is mostly insensitive to the ordering of the pointsin the database

Disadvantages of DBSCAN

Proper determination of the initial values of the parameters ε and *MinPts*is difficult For *n* data objects, without any specialstructure or spatial indexing, the computational complexity is O(n2); with spatial indexing it isO(n log n).[11]

4. OPTICS

While the partitioning density-based clustering algorithm DBSCAN can only identify a—flat clustering, the newer algorithm OPTICS computes an ordering of the points augmented by additional information, i.e. the reachabilitydistance, representing the intrinsic hierarchical (nested) cluster structure.[11] ***Core-distance*:** Let *p* be an object from a database DB, let *Nε(p)*be the ε-neighborhood of *p*, let MinPts be a natural number and let MinPts-dist*(p)* be the distance of *p* to its MinPts-the neighbor. Then, the *core-distance* of *p*, denoted as core-dist *ε,* MinPts*(p)*is defined as MinPts- is t*(p)* if|*Nε(p)* |≥ MinPts and INFINITY otherwise. ***Reachability-distance*:** Let *p* and *o* be objects from a database DB, let *N ε(o)* be the ε-neighborhood of *o*, let dist*(o, p)* be the distance between *o* and *p*, and let MinPts be a natural number. Then the *reachability distance* of *p* w.r.t. *o* as shown in fig.5,denoted as reachability-dist *ε,* MinPts*(p,o)*, is defined as max*(*core-dist *ε,* MinPts*(o),* dist*(o,p))*.[13]

The OPTICS algorithm creates an ordering of a database, along with a reachability-value for each object. Its main data structure is a *seed list*, containing tuples of points and reachability-distances. The seed list is organized ascending reachability-distances. Initially the seed list is empty and all points are marked as *not-done*.[12]

Advantages

•It solves the problem of finding good clusters if data has changeable density

• It outcomes the objects in a particular ordering.

Disadvantages

• It expects some kind of density decline to find cluster borders.

• It is less sensitive to erroneous data.

5. DENCLUE (Density based clustering) Main concepts are used here i.e. influence and density functions. Influence of each data point can be modeled as mathematical function and resulting function is called Influence Function. Influence function describes the impact of data point within its neighborhood. Second factor is Density function which is sum of influence of all data points. According to DENCLUE two types of clusters are defined i.e. Centre defined and multi Centre defined clusters.[13].

In Centre defined cluster a density attractor. The influence function of a data objects y Є F is a function. Which is defined in terms of a basic influence function F, F (x) =- F (x, y). The density function is defined as the sum of the influence functions of all data points. DENCLUE also generalizes other clustering methods such as Density based clustering;

partition based clustering, hierarchical clustering.

In density based clustering DBSCAN is the example and square wave influence function is used and multicenter defined clusters are here which uses two parameter σ = Eps, ξ = MinPts. In partition based clustering example of k-means clustering is taken where Gaussian

ISBN 978-81-939386-1-4

Influence function is discussed. Here in center defined clusters ξ=0 is taken and σ is determined. In hierarchical clustering center defined clusters hierarchy is formed for different value of σ.

5.1 Algorithm

- Take Data set in Grid whose each side is of 2σ
- Find highly densed cells i.e. Find out the mean of highly populated cells If d(mean(cl), mean(c2)) < 4a then two cubes are connected.
- Now highly populated or cubes that are connected to highly populated cells will be considered in determining clusters.
- Find Density Attractors using a Hill Climbing procedure.
- Randomly pick point r. Compute Local 4 σ density Pick another point (r+1) close to previous computed density. If den(r)< den(r+1) climb.
- Put points within (σ /2) of path into cluster.
- Connect the density attractor based cluster.[14]

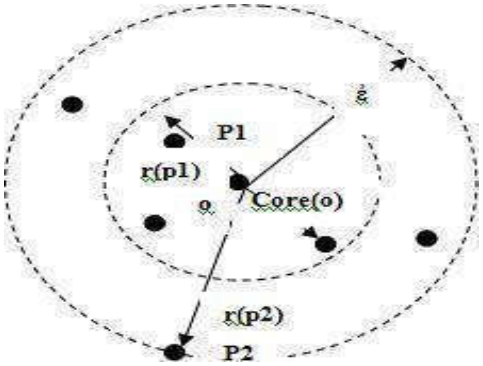


Fig.5Core-distance (o),reachability-distances r(p1,o), r(p2,o) for MinPts=4

Advantages:

- It detects erroneous data very well.
- It allows a brief description of non-spherical shaped clusters in high-dimensional data sets.
- Its processing is much faster than DBSCAN.

Disadvantages

• It needs many constants.

• It is less sensitive to outliers

Table.1 Comparison of Density-Based Clustering Methods[15]

Name	Noise	Varied Density	Primary Input Required	Complexity	Data Type	Cluster Type	Data Set
DBSCAN	Yes	No	Cluster radius, Minimum no.of Objects	O(nlogn)	Numerical	arbitrary	High-Dimensional
OPTICS	Yes	Yes	Density Threshold	O(nlogn)	Numerical	arbitrary	High-Dimensional
DENCLUE	Yes	Yes	Radius	$O(n^2$	Numerical	arbitrary	High-Dimensional

5. Conclusion

Clustering algorithms are attractive for the task of class identificationinspatialdatabases.In this paper, several density based clustering algorithms proposed for clustering have been discussed with its advantages and disadvantages. And also it gives a comparative analysis of three density-based clustering algorithms i.e. DBSCAN, OPTICS and DENCLUE.

References

[1] RashiChauhan, PoojaBatra, ”A Survey of Density Based Clustering Algorithms”IJCST Vol. 5, ISSue 2, April - June 2014

[2] PoojaBatraNagpal ,”Comparative Study of DensitybasedClusteringAlgorithms”, I*nternational Journal of Computer Applications (0975 – 8887)Volume 27– No.11, August 2011*

[3] International Journal of Computer Applications (0975 – 8887) Volume 27– No.11, August 2011 44 Comparative Study of Density based Clustering Algorithms PoojaBatraNagpal Department Of Computer Science NIT KurukshetraPriyankaAhlawat Mann Department Of Computer Science NIT Kurukshetra

[4] Fast density clustering strategies based on the *k*-means algorithm, Elsevier,Volume 71, November 2017, Pages 375-386

[5] B.G. Obula Reddy, Dr. MaligelaUssenaiah,"Literature Survey On Clustering Techniques", IOSR Journal of Computer Engineering, Vol. 3, Issue 1, July 2012

[6]HenrikBäcklund, Anders Hedblom, NiklasNeijman, "DBSCAN A Density-Based Spatial Clustering of Application with Noise", 2011

[7] Density-ratio based clustering for discovering clusters with varying densities,pattern recognition,Elsevier,Volume 60, December 2016, Pages 983-997

[8] Yong-Feng Zhou, Qing-Bao Liu, S. Deng, Q. Yang, An Incremental Outlier Factor Based Clustering Algorithm, Proceedings of First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 Nov 2002

[9] [DBCAN] Ester M., Kriegel H.-P., Sander J., Xu X.: "A DensityBased Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, 1996, pp.226-231.

[10]S.Chakraborty,Prof.N.K.Nagwani,*Analysis andStudyofIncremental DBSCAN Clustering Algorithm*, International Journal of Enterprise Computing And Business Systems, Vol. 1, July2011

[11] [OPTICS] Ankerst, M., Breunig, M., Kreigel, H.-P., and Sander, J. 1999. OPTICS: Ordering points to identify clustering structure.In Proceedings of the ACM SIGMOD Conference, 49-60, Philadelphia, PA.

[12] M. Ankerst, M. M. Breunig, H. P. Kriegel and J. Sander, OPTICS: Ordering Points To Identify Clustering Structure, at International Conference on Management of Data, Philadelphia, ACM 1999

[13] AasthaJoshiet.al.,"AReview: Comparative Study of Various Clustering Techniques in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 3, March 2013.

[14] Hinneburg and D. Keim, "An efficient approach to clustering Large multimedia databases with noise," in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD"98), 1998, pp. 58–65.

[15] An Empirical Evaluation of Density-Based Clustering Techniques Glory H. Shah, C. K. Bhensdadia, Amit P. Ganatra, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Design of Chaotic Behavior for programmable cellular automata based Symmetric Key Encryption Algorithm

Shanthi S
Malla Reddy College of Engineering and Technology Hyderabad, Telangana State, India
shanu_shivak@yahoo.com

Chandrasekar V
Malla Reddy College of Engineering and Technology Hyderabad, Telangana State, India
*drchandru86@gmail.com**

Abstract

Cellular automata are highly parallel and distributed systems which are able to perform complex computations. Cryptographic techniques are very important in these times dominated by the growth of digital information storage and transmission. CA are an attractive approach for cryptographic applications. They are simple, modular logic systems that can generate good quality pseudorandom bit streams as required in robust cryptographic systems. Further advantage is that they can be easily and efficiently implemented. This paper focuses on using cellular automata in cryptography, thereby bringing the advantages of using cellular automata in cryptography.

Keywords: Cellular automata, types of cellular automata, Cryptography, Advantages of using Cellular automata in cryptography.

1. Introduction

In this age of information, communications and electronic connectivity, security is a topic of general interest that should never be underestimated. The security of data bases, of data communications, of Internet connections, of scientific research and of personal e-mail and phone calls are some examples where the encryption of data/information plays a major role. Therefore, cryptography has become an important field of research in theory and applications development.

Because of its importance, cryptography is nowadays a science by itself, strongly related to other modern research fields as complexity theory, chaos, dynamical systems, computing theory etc. The state-of-the art for the field of cryptography is probably classified as it has military applications, but for the public domain a good reference can be found in [1] and [2]. The *encryption* of a message/data file/other information is a process (algorithm) that modifies this message/data file/information making it completely unintelligible, except for the person who knows the encryption *key*. The *key* refers to the encryption algorithm that has been used - in fact, to the reverse algorithm that should be used for decryption - and the particular parameters that have been used during the encryption. The *decryption* algorithm should render the original message/file/information complete and unaltered. Encryption can be achieved by constructing two different types of ciphers—stream ciphers and block ciphers. A block cipher is one in which a message is broken into successive blocks that are encrypted using a single key or multiple keys. In a stream cipher the message is broken into successive bits or characters and then the string of characters is encrypted using a key stream. The *cryptographic scheme* refers to the assembly of encryption and decryption algorithms.

An ideal cryptographic scheme or algorithm has not been developed yet, as an ideal cryptographic scheme implies:

- no data expansion during encoding process;
- fast encoding algorithm;
- small dimension key;
- fast decoding algorithm;
- correct and complete rendering of message after encryption/decryption;
- invulnerability to attacks.

The last point is a major issue in cryptography; complex mathematical studies and research have to be done in order to establish the vulnerability of each cryptographic scheme. In simple words, this answers the basic question: how difficult is to break the code? This “difficulty” has to be established in terms of complexity, cost and computing time. Therefore, depending of the particular applications, sometimes it is enough to have a code and cryptographic scheme that requires a long search for the key, although the process is very simple. This is the situation for the briefcases with cipher, where the breaking process is quite simple: one has to try *all* possible numbers in order to find the right one. Cellular automata are applied with success in cryptography mainly because their vast phenomenology and apparently big complexity require a very long computing time to break well-chosen cryptographic schemes. There are indeed a lot of parameters and factors that can drastically affect the encrypted message (cyphertext) and therefore the complexity of the attack is considerably increased. Cellular automata offer an ideal mathematical model for massive parallel computation, but most research and applications in cellular automata domain are done through simulation. However, it is obvious that only the hardware implementations of this model fully exploit its computing and high-speed possibilities. In particular, cellular automata applications in cryptography are efficient because of the massive parallelism of the model. When implemented by means of other computing systems (simulated in software or emulated with microcontrollers etc.) the parallel processes are in fact executed sequentially. Special cellular automata hardware is the only means to benefit of all the advantages of the model.

ISBN 978-81-939386-1-4

2. Cryptography with Cellular Automata

2.1. The Cellular Automata Model

Basically, cellular automata are parallel systems that consist of a typically large number of finite automata (finite state machines) as elementary "cells". The cells are locally connected, in other words the global network supports only local connections. The system evolves through local changes: all cells are updated synchronously, depending on their own current state and on the states of the neighbouring cells in the network. The computing task performed by cellular automata is generally conceived as the global evolution of its configuration, starting from the initial configuration (input data) and leading to an intermediate or final configuration or attraction cycle that are interpreted as a result. This almost "visual" perceptible computation is quite an advantage in task like modeling and simulation, image processing and cryptography.

The Cellular Automata formalisms [Wol86] are well suited to describe some kind of real complex systems with different description levels.

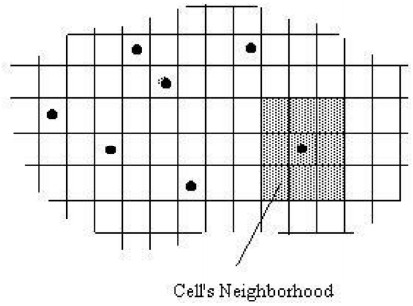


Figure 1 Sketch of a Cellular Automata

2.1.1 The three fundamental properties of CA

1. Parallelism: A system is said to be parallel when its constituents evolve simultaneously and independently. In that case cells updates are performed independently of each other.

2. *Locality:* The new state of a cell only depends on its actual state and on the neighborhood.

3. *Homogeneity:* The laws are universal, that's to say common to the whole space of CA.

Two common two-dimensional neighborhoods are the von Neumann neighborhood, in which each cell has neighbors to the north, south, east and west; and the Moore neighborhood, which adds the diagonal cells to the northeast, southeast, southwest and northwest2.

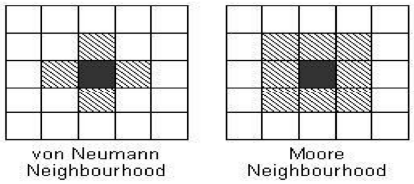


Figure 2 Types of Neighbourhoods

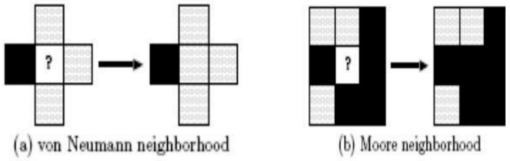


Figure 3 Two examples of a cell with a one-step update. The center cell examines its neighborhood and applies the update rule

2.1.2 Basic Rules of CA

Rule 1: Survival – a live cell with exactly two or three neighbors stay alive[7][8]

Rule 2: Birth – a dead cell with exactly three live neighbors becomes alive

Rule 3: Death – a cell dies due to 'loneliness' if it has only one neighbor or due to 'crowding' if it has more than four neighbors

Cellular automata are defined by the following elements:

- topology and dimension of the lattice of cells;
- the set of the neighbours of each cell that are involved in the next state's computation;
- the number of states of each cell (identical for all cells; binary automata, for instance, have only two states per cell: 0/1, visually translated as white/black);
- the local rule that gives the next state of the cells.

In cryptography the main topologies that are applied are liner and two-dimensional, referring to a row or matrix of "cells".

2.2 Characteristics of cellular automata.

Cellular automata were first introduced by John Von Neumann (1966), as a means of modeling the nature of self-reproduction in biological systems. The Game of Life constructed by Conway constitutes a famous example of cellular automata which may be easily simulated on a personal computer. Cellular automata are discrete dynamical systems in which the space, time, and states are discrete. Despite their conceptual simplicity, cellular automata may reveal very complex behavior. A cellular automaton is a mechanism for modeling systems with local inter-actions. It consists of a regular array of identically programmed units called cells which interact with their neighbors subject to a ¯finite set of prescribed rules for local transitions. All cells form a regular spatial lattice. Time progresses in discrete steps. The state of a cell at time t + 1 is a function only of its own state and of the states of its neighbors at time t. All cells states are updated synchronously. In order to establish a rigorous mathematical definition, we need some basic notions. Cellular automata remain very interesting for systems with high complexity. A cellular automaton is a mathematical model which is perfectly suited to complex systems containing a large number of simple identical components with local interactions. Cellular automata have also been considered as a means for symbolic dynamics. Such an approach offers many advantages. In theory, it is a

ISBN 978-81-939386-1-4

convenient method to represent many discrete or continuous processes conventionally described by partial differential equations. In practice, cellular automata may be viewed as parallel-processing computers of simple construction. Because of this, Cellular automata have been used to study complex systems widely.

2.3. Complexity of Cellular Automata

The cellular automata model is inspired from the natural model of complex systems that often consist of a large number of simple basic elements, having only local interactions that lead to a complex global behaviour. This model is an important research and simulation tool in the science of complexity. Strictly speaking, the cellular automata model and also the cellular automata evolution are not complex, as its structure and rules are very regular and simple. But the huge dimension of the rules and configuration spaces confers to its phenomenology a considerably great *apparent* complexity. In the quite simple example of binary linear cellular automata with 100 cells (a modest dimension) with local rules involving the central cell and two neighbours on each side (binary functions with 5 variables), there are: 2^{100} global configurations, which is approximatively 10^{30} and 2^{32} possible local functions, which by a rough approximation means around 4,000,000 functions.

Not all possible functions are of practical interest, but even in this conditions a search in the local functions' space is a very long computing task. This is related to the difficulty of cellular automata synthesis: there is no algorithm that gives the appropriate local function for a specific application. The universality of the cellular automata model is theoretically proven, but the practical applications are still waiting for development tools, since the experiment is, by now, the main means of cellular automata synthesis. The work of S. Wolfram [7] imposed a certain order in the space of local rules of functions, mainly for linear cellular automata, dividing it into four complexity classes (that express how complex is the evolution of the system govern by a certain rule). Class 3 of Wolfram's classification contains the automata that have an apparently chaotic evolution, also strongly depending on the initial global configuration. Such an evolution is illustrated in Fig. 1, where time is on the vertical axis, downwards, and the horizontal bit string is the configuration of a binary linear cellular automata. The *class 3 automata* are ideal for applications like random sequences generation and cryptography.

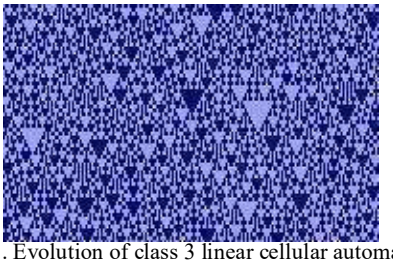


Fig. 1. Evolution of class 3 linear cellular automata

2.3.1 Chaos and complexity

Wolfram gives a rough geometrical analogy of behavior of these four classes:

1. **Class 1** - limit points
2. **Class 2** - limit cycle
3. **Class 3** - chaotic - "strange" attractor
4. **Class 4** - more complex behavior, but **c**apable of universal computation [4][5]

2.3.1.1 Class 1 cellular automata

After a finite number of time-steps, class one automata tend to achieve a *unique state* from (nearly) all possible starting conditions.

2.3.1.2 Class 2 cellular automata

This type of automata usually creates patterns that *repeat periodically* (typical with small periods) or are stable. One can understand this type of CA's as a kind of *filter*, which makes them interesting for digital image processing

2.3.1.3 Class 3 cellular automata

From nearly all starting conditions, this type of CA's lead to aperiodic - chaotic patterns. The statistical properties of these patterns and the statistical properties of the starting patterns are almost identical (after a sufficient period of time). The patterns created by this type of CA's (usually one dimensional CA's) are a kind of self-similar fractal curves.

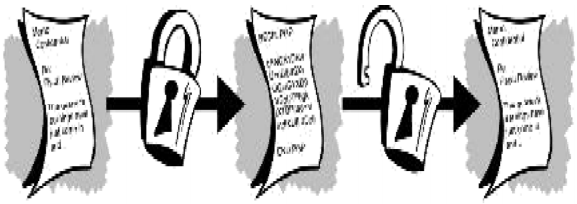
2.3.1.4 Class 4 cellular automata

After finite steps of time, this type of CA's usually "dies" - the state of all cells becomes zero. Nevertheless a few stable (periodic) figures are possible. One popular example of an automaton of this type is the Game of Life. In addition to that Class 4 automata can perform universal computation This class of CA's show a high irreversibility in their time development.

The first three types can be read as *Cantor* sets with a certain dimensionality, either in countable or in *fractal* dimension. Class 3 is the most frequent class. With increasing k and r the probability to find class 3 automaton for an arbitrary selected rule is again increasing.

3.Cryptography

Encryption is the science of changing data so that it is unrecognizable and useless to an unauthorized person. Decryption is changing it back to its original form. The most secure techniques use a mathematical algorithm and a variable value known as a 'key'. The selected key (often any random character string) is input on encryption and is integral to the changing of the data. The EXACT same key MUST be input to enable decryption of the data. This is the basis of the protection.... if the key (sometimes called a password) is only known by authorized individual(s), the data cannot be exposed to other parties. Only those who know the key can decrypt it. This is known as 'private key' cryptography, which is the most well known form.



Plaintext encryption cipher text decryption plain text

Figure 4 Cryptography basics

Data that can be read and understood without any special measures is called plaintext or cleartext. The method of disguising plaintext in such a way as to hide its substance is called encryption. Encrypting plaintext results in unreadable gibberish called cipher text. You use encryption to make sure that information is hidden from anyone for whom it is not intended, even those who can see the encrypted data. The process of reverting cipher text to its original plaintext is called decryption.

4. Cellular Automata in Cryptography

The very large phenomenology of the cellular automata model and its apparently big complexity offer a good basis for applications in cryptography (cellular automata are not the only dynamical systems applied in cryptography, and some of the basic principles of cryptography with cellular automata also stand for other dynamical systems). Massive parallelism is another feature of cellular automata that make this model attractive for cryptography, since lot of computation is often necessary in real-time applications. However, this parallelism, when emulated in software or in sequential hardware, disappears. In the last two decades many results have been obtained in this direction. A good review of some cryptographic schemes and systems with cellular automata proposed in the scientific literature is given in [5] and [6]. In the next section some of these will be discussed as possible basis for hardware implementations, taking as valuable the theoretical results concerning the efficiency and invulnerability of the schemes.

5. Other Advantages of Cellular Automata

The design style under VLSI technology prefers the simple, modular and local connected logic circuit structure. Cellular automata are ideal from the hardware designer's point of view. The local connectivity, regularity and the simple basic components make CA very appropriate to implement low-cost and robust massive parallel machines or application-oriented circuits. The reason why there are not so many cellular-automata inspired electronic circuits is mainly the difficulty of the synthesis for this computing model; however, dedicated circuits for specific applications (signal generators, associative memories, image pre-processing blocks, various simulators for natural phenomena) have been successfully designed and produced [6].

Apart from cryptography cellular automata can be used in variety of applications because of its simpler structure providing solution to the complex problems.

6. Conclusion

This paper has introduced the basic concepts of cellular automata and cryptography and some applications of cellular automata in cryptography. cellular automata along with cryptography can be used to model a wide variety of real world applications.

7. References

[1] William Stallings, *Cryptography and Network Security* Prentice-Hall 2003 New Jersey, USA
[2] Brian A. LaMacchia, *.NET Framework Security* Addison Wesley 2002, USA [1]Toffoli, T., & Margolus, N. (1987). Cellular Automata Machines. MIT Press.
[3] S.Wolfram, *Cellular Automata and Complexity*, Addison-Wesley Publishing Company, 1994
[4] J. Kari "Reversibility of 2D Cellular Automata is Undecideble', *Physica D, Vol 45:379-385,* 1990
[5] H.Gutowitz *Method and Apparatus for Encryption, Decryption and Authentification using Dynamical Systems* US Patent 5365589, 1994
[6] P. Chaudhuri, D. Chowdhury, S. Nandi, S. Chattopadhyay, *Additive Cellular Automata: Theory and Applications*, IEEE Computer Society Press, 1997.
[7] Wolfram, S. (2002). A New Kind of Science. Wolfram Media, Inc., Champaign, IL.
[8]1986. Wolfram, S. (Ed.). *Theory and Applications of Cellular Automata*. Singapore: World Scientific.
[8]V. Vemuri, 1978, *modeling of Complex Systems: An introduction.* Chapter 3 pp. 67-88
[11]Pidd, M. 1992a. *Computer simulation in management science* 3rd ed, Chichester.: John Wiley & Sons.

ISBN 978-81-939386-1-4

IoT BASED SMART VILLAGE FOR THE RURAL DEVELOPMENT

Dr.J.Gladson Maria Britto,Professor/CSE, Dr.V.Bhoopathy, Professor/CSE

Malla Reddy College of Engineering,Maissamaguda

Abstract:

In this paper, IoT based smart village system is developed to support value-added services for various attributes of the village and for the people, while it still being a broad and complex category that are characterized by specific application domain. Rural developments are designed to support the Smart village mission, which aims at exploiting the most advanced communication technologies. The global focus on waste, energy and water management and conservation and the cloud based system plays a key role in extending the connected benefits of the smart village beyond the distribution, automation and monitoring being done by utility. IoT based Monitoring system will help consumers to monitor their own usage and adjust behaviours. The proposed systems will eventually regulate automatically by operating during off-peak energy hours and connect to sensors to monitor occupancy, lighting conditions, and also optimized irrigation management for those attributes are incorporated. This paper will address and discuss the technical solutions for the energy management and smart irrigation system which can be adopted in the rural development mission

Keywords: Cloud system, Energy management, IoT, Smart village, Smart irrigation system, Demand response (DR)

1. INTRODUCTION

There are huge challenges in realization of a rural development that monitors and integrates all of the village infrastructure and services to leverage the collective intelligence. The development of a IoT based smart village includes Cloud based network which can provide a virtual infrastructure to process and integrate the analysis tools monitoring equipment, storage, and visualization platform within the system.IT- OT convergence which intend to smart billing and data analytics in energy management. Waste collection system enhanced with cloud based IoT services which enable dynamic scheduling and routing in a waste collection system seems to be an efficient system.

The access to sustainable energy services acts as a basic catalyst for smart village development extending the provision of efficient management of energy, water, waste and various other attributes. Major parts of rural areas are not stand-alone system it remains a part of a cluster, which are related to each other. These clusters explains the potential for development and it possess economic drivers, derive

Location and competitive advantages. Urban is named after the development of these clusters which aims to strengthen the rural areas by provisioning of physical infrastructure, economic and social facilities. The envisaged
components in each cluster are listed below:

- ✓ Education system.
- ✓ Sanitation
- ✓ Waste management.
- ✓ Road system Inter village connectivity.
- ✓ Lighting control
- ✓ Fleet management
- ✓ Digital Literacy and People Service Centres.
- ✓ Economic activity based Skill development programme
- ✓ Agro Processing, Agri Services, Storage and development.

IoT based smart village can help to reduce cost through improved process efficiency, asset utilization

and productivity. The tracking of devices is improved using sensors and communication devices which can be benefited from real-time datas and analytics help them make smarter decisions. The growth and convergence of data, processes and things on the internet would make such connections more relevant and important, creating more opportunities for people, businesses and industries.

Represents the overall block diagram of the smart

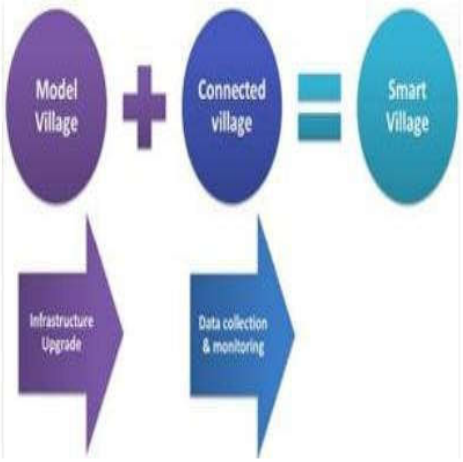


Figure 2 Smart village transition diagram

village system. The enormous IT infrastructure is required by rural development along with the huge financial support which is to be incorporated. Sensors, thousands of networking equipment and computing devices are built in this complex network. Operational and maintenance cost of such a complex real time system will be much higher which is evident to meet stringent reliability and efficiency improvement. In case of smart irrigation management system each field has to be fitted with a sensors and data control unit which are highly efficient and reliable.



Figure 1 Overall block diagram

NEED FOR SMART VILLAGE

Each country has developed a reputation as a global leader in upgrading their city as smart city initiatives in its larger urban areas. The Rural areas are in need of essential infrastructure like roads, drinking water and power. The future development mainly concentrates on improving big metropolises into connected cities but failed to see where most of the population resides. Villages more than cities need to be made smart for the overall improvement and development of the country. The Development of opportunities for youths in villages, thereby discouraging migration to cities. Farming remunerate occupation, with guidance and mentoring to farmers on how to get the best yield and market at remunerative prices for the future rural development. Proper implementation which presides over the benefits such as crop insurance, soil health card, and pesticides which can reach the grassroots. Fig.2 explains the smart village transformation.

The paramount importance should be given to develop an economically viable and culturally sensitive ecosystem in villages. The challenges remain the same, direct access to the global market has been a major challenge largely due to multiple intermediaries and lack of skilled workforce. The large population lives in villages, we always fail to improve economic potential and basic services by creating smart village.

ISBN 978-81-939386-1-4

The most villages lack essential infrastructure like

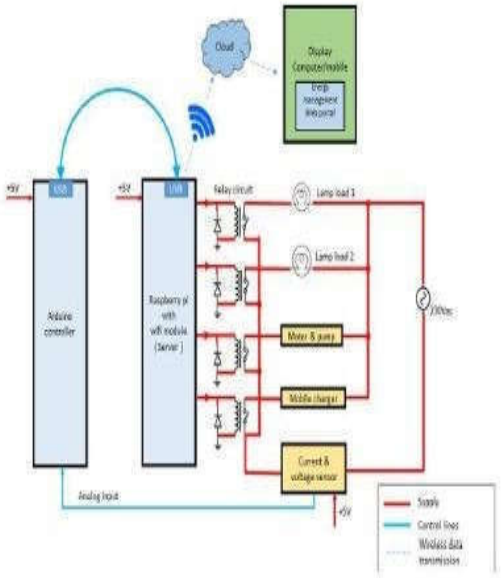


Figure 3 Circuit diagram of energy management module

proper irrigation system, electricity and water. To overcome this challenge, three strategy can be followed:

Provide education on technology that supplements indigenous skills,
Ensure digital and IT awareness, and
Connect skilling-interventions to market.

Smart Village development enable and access to sustainable energy services which paves way for a development – enabling the provision of good education and healthcare, optimized energy management, sanitation, avoid wastage of water, and enhanced security, gender equality and the growth of productive enterprises to boost incomes, democratic engagement.

LACK OF AWARENESS Almost with all Indian villages, lack of awareness about schemes and subsidies in various areas by the government, is the major problem they face. This has created a major hindrance on the occasions when we tried to update technologies and implement some projects like- biogas plant installation, Iot based installation and many more. An unhealthy wariness and lack of awareness about banking facilities still exists amongst the rural people. In order to incorporate any such advancement or technologies first step is to create awareness among the people.

2.ENERGY MANAGEMENT The challenges in load control for the power grid is now more severe than ever, due to advancement in communication layer and the creation of a two-way infrastructure for real time communication between people and the utility. The operator of the smart grid access the information and communication technologies to enhance grid security and reliability. They enforce controllable use of energy, and incorporate various components such as green resources, distributed generator and power storage premises. Fig. 3 explains the energy management circuit diagram

Energy management is the process of monitoring, controlling, and conserving energy in a building or organization. It can also be defined as the strategy of adjusting and optimizing energy, using systems and procedures so as to reduce energy requirements

Typically, this involves the following steps:

Metering your energy consumption and collecting the data.

Finding opportunities to save energy, and estimating how much energy each opportunity could save. You would typically analyze your meter data to find and quantify routine energy waste, and you might also investigate the energy savings that you could make

by replacing equipment (e.g. lighting) or by upgrading your building's insulation.
Taking action to target the opportunities to save energy. Tracking your progress by analysing your meter data to see how well your energy-saving efforts have worked.

The Demand response (DR) scheme shows that incentives are economically balanced according to the effective change of consumer behaviour. It includes minimising or shifting consumption, and using standby generation to change of electricity use from the grid to onsite generation. Notice periods of about 2-4 hours was considered by the operators as switching time to safely implement a curtail plan with minimal impact on machineries and accessories.
There are various demand response as follows

i) time of use (ToU)
ii)Critical Peak Pricing
(CPP)
iii) and Critical Peak Rebate (CPR)

Involvement in DR programs enhance the system operators with an intimation of details of the customer which includes resource size which is willing to curtail or shift load according to the response. Since the participation is voluntary in some of these DR programs can comes under the direct load control and hence utility often does not have physical control of the customer's load response. Incentive based demand side management is been followed in the smart rural development to reduce the peak demand, tariffs and minimize the outages.

The various proposed Demand response techniques involves in peak clipping or tariff based technique which directly cuts off or shut down the loads. This project proposes an idea which is a combination of both load shifting and tariff based. Here, the consumer is notified during the peak demand period and recommended to shift the interruptible loads to off peak hour.

Opinion of customer is also asked whether the loads can be shifted, if he opts then the tariff is increased several times than normal tariff. Hence the consumer will not be willing to use the interruptible loads like chargers, washing machines at an increased tariff. Obviously the peak demand decreases and hence ensures balancing of supply and demand. The fig shows the IoT based network for energy management. the energy monitoring screen which is connected to the cloud. Energy management is also concerned with balancing of supply with the demand.
It is classified into two categories namely

i)Source side management(SSM) and
ii)Load or demand side
management(DSM)

The ultimate aim of all these techniques is optimal balancing of supply with demand so that undesired blackouts and outages are eliminated. For successful implementation of demand response technique, modern equipment has to be built in the residential sector. This is achieved with the help of smart meters and smart grids with a daylong monitoring at the control Centre. The addition of a communication layer to existing grid makes it a smart grid. The smart meter measures various parameters and also help in saving the energy consumed for an hour or less.

The IEEE Smart Village network is involved in the design and monitoring of attributes of smart village

which improves technology innovation and involvement. This is mainly important when intimating when the device may need to adapt, because power-system configurations might evolve over time. Best example is when multiple solar based home systems wired together to build a rural scale

ISBN 978-81-939386-1-4



micro grid. Smart metering and billing are additional requirements to recover expenses needed for installation, monitoring, and maintenance. Renewable energy sources plays a vital role in rural electrification.

3. SMART IRRIGATION SYSTEM

The agricultural productivity is dependent on efficient irrigation system. When water is spread over a wide area covering all the places and temporally distributed it matches with field water demand hence efficiency is achieved.85% of the fresh water is consumed by the agriculture and it remains dominant due to increase in population and the increasing of demand for food].Field of agriculture has seen the rapid advancement.

Irrigation system is improved by latest technologies. As precision advanced irrigation scheduling plays an important role in reduction in water wastage. The usage of various monitoring and controlled system is increased by people in order to increase the yield

Soil moisture plays a key role in the life of the plant. Nutrients in the soil solution provide the plant with the food it needs to grow. Water is also essential for regulating plant temperature through the process of transpiration. Plant root systems are better developed when growing in moist soil. Excessive levels of soil moisture, however, can lead to anaerobic conditions that can promote the growth of plant and soil pathogens. Agricultural production is affected by poor irrigation management, it is necessary to develop strategies to optimize irrigation. The monitoring and control of various factors derived from field such as humidity, water level, temperature, and human interaction is governed by automated irrigation system. This system consists of controllers and a wireless sensor infrastructure used to transmit the detected values. The yield can be improved by various automated system which can help the farmers. Sensors can be placed anywhere in the field and system is also tested for different temperature. Only minimum deviation is observed in the sensor output. In recent years communications Infrastructure in the IoT is improved rapidly which can meet the demands between the physical world "things" and "human". Thus, the remote objects are controlled using the smart gadgets. Fig shows the drip irrigation system which is involved the smart irrigation network.

Fig 4. shows the drip irrigation system

Drip irrigation is a form of irrigation that saves water and fertilizer by allowing water to drip slowly to the roots of many different plants, either onto the soil surface or directly onto the root zone, through a network of valves, pipes, tubing, and emitters.

1. Salient Features

50 to 65% water saving compared to control method.

Crop attains early maturity.
Crop quality and yield
increased. Requires least land
levelling.
Poor quality water can be used.

2. Advantages of proposed drip irrigation system

High quality and increased fruit
size Suitable for all types of soil
Easy method of fertigation
Saving in labour and field preparation cost

Water is provided directly to the roots of the crop through drip irrigation through which wastage of water is minimized and water resources are

optimized. This system is benefited by farmers which will increase the yield of the crop. This smart drip irrigation system remains as a useful system as it automates and manages the watering without any human intervention. Depending upon the weather conditions the system is managed and automated response is generated. The water flow is managed using the relay controlled valves. Fig.7 shows the irrigation monitoring screen.

Conclusion:

The aim of all these techniques in energy management is optimal balancing of supply with demand in which undesired blackouts and outages are eradicated. For successful implementation of demand response technique, modern equipment has to be implemented the rural areas. The automation in irrigation will be ensured at low cost and high accuracy which minimize the water consumption.

This paper meet out many facets of rural development and design which includes energy, environmental, economic impact using various technologies.

References

[1] Marceau, Jane. "Introduction: Innovation in the city and innovative cities". Innovation: Management, Policy & Practice 10.2-3 (2008): 136-145.

[2] A. P. Castellani, M. Dissegna, N. Bui, and M. Zorzi, "Web IoT: A web application framework for the internet of things," in Proc. IEEE Wireless Commun. Netw. Conf. Workshops, Paris, France, 2012.

[3] R. Bonetto, N. Bui, V. Lakkundi, A. Olivereau, A. Serbanati, and M. Rossi, "Secure communication for smart IoT Objects: Protocol stacks, use cases and practical examples," in Proc. IEEE IoT-SoS, San Francisco, CA, USA, 2012, pp. 1–7.

[4] P. Suresh, "Understanding Challenges in e-Governance," Better Account with egovernance, pp. 61–63, 2011.

[5] N. Bressan, L. Bazzaco, N. Bui, P. Casari, L. Vangelista, and M. Zorzi, "The deployment of a smart monitoring system using wireless sensor and actuator networks," in Proc. IEEE Smart Grid. Comm., Gaithersburg, MD, 2010, pp. 49–54.

[6] Alghamdi, Ibrahim A., Robert Goodwin, and Giselle Rampersad. "E-government readiness assessment for government organizations in developing countries." Computer and Information Science 4.3 (2011): p3.

[7] International Journal of Information and Education Technology, Vol. 6, No. 5, May 2016 ICT Based Education for Literacy in India: Vision Beyond 2015.

[8] Fuller, J. C., Schneider, K. P., and Chassin, D. (2011). Analysis of residential demand response and double-auction markets. In IEEE Power and Energy Society General Meeting.

[9] Loganthiran, T., Srinivasan, D., and Shun, T. Z. (2012). Demand side management in smart grid using heuristic optimisation. IEEE Transactions on Smart Grid, 3(3), 1244– 1252.

[10] M. Kovatsch, S.Duquennoy and A. Dunkels, "A Low- Power CoAP for Contiki ", In Proceedings of the IEEE Workshop on Internet of Things Technology and Architectures, Valencia, Spain, October 2011.

[11] R.Hussain, J.Sehgal, A.Gangwar, M.Riyag " Control of irrigation automatically by using wireless sensor network" International journal of soft computing and engineering, vol.3, issue 1, march 2013, pp.48t 324-328.

[12] Fan TongKe, "Smart Agriculture Based on Cloud Computing and IOT", Journal of Convergence Information Technology, Vol. 8, Issue 2, pp.1, Jan 2013.

[13] Hong, Insung, Sunghoi Park, Beomseok Lee, Jaekeun Lee, Daebeom Jeong, and Sehyun Park.

T. Anagnostopoulos, A. Zaslavsky, "Effective Waste Collection with Shortest Path SemiStatic and Dynamic Routing", IEEE 14th International Conference on Next Generation Wired/Wireless

ISBN 978-81-939386-1-4

Evaluating Reliable Node in Wireless Sensor Network using ERLEN model

M.S.Nidhya[1]
Assistant Professor,
Dept of Computer Science,
VISTAS,
Chennai,TamilNadu.
{nidhyaphd@gmail.com}

Pr Dr.A.Kannagi [2]
Professor,
Dept of Computer Science and Engineering,
Malla Reddy College of Engineering,
Secunderabad, Telangana State.
{kannagianbu@gmail.com}

Dr. R.Chinnaiyan[3]
Associate Professor, Dept of Computer Application,
CMR Institute of Technology,
Bangalore,Karnataka.
{vijaychinns@gmail.com}

Abstract-In Wireless Sensor Networks (WSN), many sensors will be scattered in the environment. The main objective of the sensors is to collect the information and pass the information to the sink node. From the sender to sink node the information/packet will be forwarded by the number of intermediate nodes. If the intermediate node is malicious or dead, an information/message will not be passed or forwarded to the sink node. In this paper, we propose ERELN: Evaluating Reliable Node in a network by node's characteristics and neighboring nodes commend. The proposed new technique will find the reliable node for packet routing and the other malicious /compromised node will be eliminated automatically from the packet routing.

Keywords: Wireless Sensor Network, Reliable node, compromised node, malicious node.

1 Introduction

Wireless sensor network consists of large number of sensors, distributed in an environment to monitor environmental conditions such as habitat monitoring, temperature, forest fire, earthquake, flood etc. Events occurred in the environment should be monitored by the sensors and it will send the information to the sink node on time. If the intermediate node will not send the message to the sink node or it will misroute the message to somewhere, the message will not be received by the sink node on time. Problems cannot be identified and rectified by the sink node. In this paper, our ERLEN model will find out the reliable node to route the message and eliminate the malicious node from packet routing.

A decision-making process of the network is supported by trust management system of WSNs [1]. It helps the nodes of WSN to deal with doubt about the upcoming actions of other nodes. As WSNs are highly application oriented, these applications bring various security needs. Existence of WSN is dependent on the cooperative and trusting nature of its nodes. Hence the trust establishment between nodes is must.

Reliable nodes in a network are evaluated by ERELN technique. Before sending a packet to the neighboring node, every sender node should check the database whether the node is reliable or not. Reliability of a node will be calculated by the node's characteristic and commends given by the superior and inferior neighboring node.

In the existing models reliability is calculated only from the neighbors who directly connected with a node. A node maintains trust above the threshold level; it will be a trusted node. If a trusted node drops a packet or misroute the packet means its reliability will get low but the trust will not reduce below the threshold level. For ex: A trusted node maintains 88% of trust, but the given threshold level is 55% , if a trusted node drops a packet means its trust will be reduced to 78%.In this paper we evaluated reliability from nodes which is superior and also inferior in the network. Reliability of a node will be increased or decreased with time based on the available evidence through direct interaction with node or commends from other reliable nodes. Reliability can also be stated as a level of confidence can be taken out from the history of transactions.

2 Related Work

[2] Reliability is evaluated on component based software system. [3] Reliability estimated for GNU complier components. [4] Evaluated reliability using communication variable for object oriented software systems using communication variables. [5] Examines and compare the protocols which support reliability for transferring information from sensor node to sink node. [6] A trust metrics is proposed to improve reliability during sink hole attack. [7] Proposed a reliability model based on the energy factor and time delay factor of a node in wireless sensor network. [8] Different methods are proposed for modeling and managing trust, it enable a WSN to be secure and reduce the communication overhead. [9] Detects neighbor based malicious node in wireless sensor network. [11] evaluated a weighted trust for a three-layer hierarchical architecture. [12] Detects node failure and link failure and find a shortest reliable path to forward a message from sensor node to sink node. [13]proposed vigilant node to detect compromised node and compared existing model with EERN model. [14] proposed to find smallest distance for path Optimization in Wireless Sensor Network. [15] proposed to predict sink node using mining algorithm for path optimization in wireless sensor network. [16] proposed advanced search in wireless network using search engines.

ISBN 978-81-939386-1-4

3 Proposed Technique for Reliability Evaluation

In this paper, we extend our previous work proposed to find trusted node, but it is based on direct trust only [6]: It is a new approach for trust calculation based on both instant trust(direct) and oblique trust(Indirect). The instant trust calculation is based on node's performance, such as data packets and control packets forward, data rate, power consumption, reliability, etc.

In the same way, the oblique trust calculation is based on a comment given by neighboring node. To calculate reliability of a node, instant trust and oblique trust are combined using traditional weighting approach.

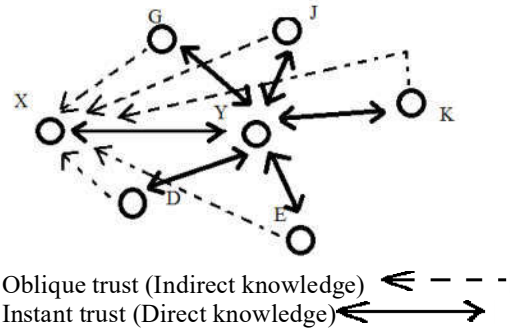


Fig-1 Reliability Evaluation of node X on Y

In the above Fig-1 node X evaluate reliability of node Y by collecting the trust from neighbor node directly (instant) and indirectly (oblique).

In our reliability calculation process, a node's trust metrics is divided into two types called superior and inferior. A function of a node is dependent on superior trust metric levels. Hence node's superior trust metrics should not below the certain prescribed level. Control packet forwarding and data packet sending should not be lower than certain prescribed level because node's functionality is hidden in these trust metrics

Reliability on any neighboring node can be evaluated from superior trust metric and inferior trust metric. A node should maintain a certain level (high) of trust in the superior trust metrics hence, product of the trust metrics are taken into an account. Similarly, all other trust metrics (of inferior), are averaged or arithmetic mean value will be taken.

To evaluate the reliability, these two (superior & inferior) are combined and called as weight to each one. Hence if a node's superior trust metric falls below the prescribed level then that node become a malicious/selfish/unreliable. Similarly, a node will become a malicious/selfish/unreliable, if the node over all instant trust is falling below a prescribed level of instant trust.

Based on trust value given by a neighboring node a reliability of a node is evaluated. Oblique trust means trust value given by directly connected node. It is calculated by collecting a trust value from all other neighboring node which is directly connected with particular node. The neighboring nodes are divided into two groups, most trusted neighbor and normal neighbor. Every node not only maintains the trust metric data but also, maintains their trust history, some neighboring nodes are most trust worthy and can be separated into most trusted/high reputation neighbors and some are into less trusted neighbors. The most trusted neighboring node must recommend positively.

If all other neighboring nodes are of lower priority, then trust value given by them will get averaged. To evaluate the oblique trust, these two are joint by giving some weight to each one. Here a node can become malicious\selfish at two different stages, one is, when any one of the superior neighboring node commends is falls below the prescribed level, and second, over all oblique trust is comes low than its prescribed level. The overall trust or total trust of any node in a network is calculated by its neighboring node.

Table 1. Trust Metric

1) Data packets dispatched
2) Control packet / message forwarded
3) Battery lifetime
4) Routing protocol performance (routing activities)
5) Reliability of informed (sensed) values/data
6) Packet address Modified
7) Availability based on beacon/ hello messages
8) Sensing Communication
9) Packet delay
10) Message cryptography.

In this reliability evaluation, trust metrics are divided into two types ,one is superior category another one is inferior category. A node's main functionality can be valued by its superior activity. Hence, node's superior trust metrics should not drop below the threshold level.

Based on the trust metrics we are taken only five parameters to evaluate reliability of the sensors. In this paper, the ten essential features of sensors are checked. These features will start from superior priority to inferior. These features are extracted from the trust metric [10].

3.1 Instant trust or Primary trust

Data collection for reliability evaluation is the one of the essential features of trust management schemes. Instant trust of a neighboring node is defined by the various trust metrics of that node in various events happened in the network. The listed trust metric for various events are vital and it is useful for trust management system of a node to take a decision. Here, depending on the application, we can hold the minimum level (threshold) to all the trust metric.

Every node should have certain prescribed level of trust in each metric exactly. Depends on the application the

superior and inferior will change. This ERLEN model considers a node only if it is having minimum threshold in each trust metric exactly. A proposed evaluation model will filter the malicious, compromised, sink hole nodes from packet routing.

Let $TM^{M}_{X,Y}$ be the set of superior trust metric on node X on node Y for different trust metrics, where M=1 to k and let $TM^{N}_{X,Y}$ be a set of inferior trust metric on node X on node Y for different trust metrics, where n=1 to l. Also W^{S}_{IN}, W^{I}_{IN} are the weights given to superior trust metrics and inferior trust metrics correspondingly to find the instant trust or primary trust. Here $W^{S}_{IN} + W^{I}_{IN} = 1$.

In this reliability evaluation model, the instant trust of any node is calculated by adding the geometric mean value of all superior trust metrics with weight any other neighboring node. Every superior trust metric must be larger than or equal to prescribed level, and arithmetic mean of all inferior trust metrics.

- Superior trust metrics calculated by geometric mean.
- Inferior trust metrics are calculated by arithmetic mean.

From this information, neighboring node instant trust is calculated and is shown below.

$$IN_{X,Y}=W^{S}_{IN}\times [\prod(R^{X,Y}_{1},R^{X,Y}_{2}\dots R^{X,Y}_{K})]^{\left[\frac{1}{K}\right]}+W^{I}_{IN}\times\frac{1}{L}[\sum(R^{X,Y}_{1},R^{X,Y}_{2}\dots R^{X,Y}_{l})]$$

$$IN_{X,Y}=W^{S}_{IN}\times[\prod_{M=1\ TO\ K}R^{X,Y}_{M}]^{\left[\frac{1}{K}\right]}+W^{I}_{IN}\times\frac{1}{L}[\sum_{N=1\ to\ l}R^{X,Y}_{N}] \quad (1)$$

3.2 Oblique trust or Secondary trust

The secondary trust of any node on any neighboring node can be evaluated from the ancillary information given by the neighboring node. In this method, the neighboring nodes are separated into most trusted and normal nodes. Trust value given by the most trusted neighboring node is calculated by the geometric mean. Applying arithmetic mean on the trust value is collected from normal nodes. Oblique trust of node X on Y is given by the below expression. Oblique trust of node X on node Y is given by the following equation.

$$OB_{X,Y}=W^{S}_{OB}\times[\prod_{I=1\ to\ r}W^{X,Si}\times T^{Si,Y}]^{\left[\frac{1}{r}\right]}+W^{I}_{OB}\times\frac{1}{s}\sum^{s}_{J=1}(W^{X,Sj}\times T^{Sj,Y}) \quad (2)$$

3.3 Selection of higher priority neighbouring node

Due to the transaction, each node of WSN should have a file to store its every neighboring node. In a WSN some nodes will be most trusted someone may not be trusted. In this research, nodes are separated as two groups, one is most trusted and another one is normal node. This separation is based on the trust value calculated from the node during last X transactions. It is a node's responsibility to verify its database to select trusted nodes if any. The ancillary information given by neighboring nodes would be applied to the geometric mean function. The ancillary information given by all other neighboring nodes would be applied to the average function in the calculation of oblique trust.

4 Reliable Node Evaluation

The reliability of a node is calculated by weight given by a neighboring node which is instant trust (IN) and oblique trust (OB). $R = W_{IN} \times IN + W_{OB} \times OB$. The W_{IN} weight given to instant trust and W_{OB} to oblique trust where $W_{IN} + W_{OB} = 1$. Weights can be allotted using different methods. Sometimes IN Instant trust will be more, and OB Oblique trust may be less i.e) $W_{IN} > W_{OB}$. Therefore a new expression will define a trust of sensor node in a network as shown below.

$$R_{X,Y} = W_{IN} \times IN_{X,Y} + W_{OB} \times OB_{X,Y} \quad (3)$$

$$R_{X,Y} = W_{IN}\left\{W^{S}_{IN}\times[\prod_{M=1\ to\ K}R^{X,Y}_{m}]^{\left[\frac{1}{K}\right]}+W^{I}_{IN}\times 1/L[\sum_{N=1\ to\ l}R^{X,Y}_{n}]\right\}+W_{OB}\times\left\{W^{S}_{OB}\times [\prod_{I=1\ to\ R}W^{X,Ni}\times T^{Ni,Y}]^{\left[\frac{1}{R}\right]}+W^{I}_{OB}\times 1/S\sum^{S}_{J=1}(W_{X,Ni}\times T^{Nj,Y})\right\}$$

Where, X is the node that is calculating reliability on node Y, K is superior trust metrics, l is inferior trust metric is a most trust worthy neighbor, s is normal neighbors.

$TM^{X,Y}_{m}$ is m^{th} superior trust metric of node X on node Y.
$TM^{X,Y}_{n}$ is n^{th} inferior trust metric of node X on node Y.
W^{S}_{IN} is superior trust metric of instant node,
W^{I}_{IN} is inferior trust metric of instant node,
W^{S}_{OB} is superior trust metric of oblique node,
W^{I}_{OB} is inferior trust metric of oblique node,

$W_{X,Ni}$ is a weight for a recommendation made by i^{th} neighbor,
$T^{Ni,Y}$ is trust given by neighbor N_i about node Y.
W_{IN}, W_{OB} are weights of instant and oblique trusts respectively.

After getting the trust value evaluate the following algorithm.

4.1 Adaptive reliability evaluation algorithm

The proposed reliability evaluation method is an adaptive trust evaluation method. First it tries to find the trustworthy neighboring node from the available trust metric database. If any node's trust value greater than or equal to prescribed level of trust then that neighboring node will be selected for packet transmission from that moment. But if no node found truthful then it finds another way for packet routing. Nodes make the transaction with the neighboring nodes and get the indirect information from all neighbors those are situated in its radio range. Then it updates the database of the node, and evaluates trust and chooses the reliable node, if a reliable node is not found then the packet is handed over to a node with high risk.

Because of first step, node's energy cannot be wasted. If a reliable neighboring node is not found within its range, then node increases its radio range to find reliable neighbor nodes through overhearing. This process is adaptive and consumes energy but it occurs in risky situations. Hence this evaluation process saves the energy of node if the trusts of neighbor node is greater than or

equal to trust threshold. This increases the trust of a node as well as life span of the entire WSN.

When a node wants to make a communication with its neighboring nodes for indirect information data, it simply sends a request command to its neighbor. This command may take only one byte in size. But, when it is getting indirect information the size of message from neighboring node may be approximately 20 to 25 bytes by assuming that each node's information takes 2 bytes and it may have 10 neighbors. The communication among the nodes for indirect information will take place only when there is a risky situation and not all the time. Hence, this adaptive ERELN method of reliable node evaluation not only energy efficient but also decrease the communication cost.

4.1.1 Process for selecting reliable node to route packet

In the below figure we clearly explained the process of reliability evaluation on a Wireless sensor network

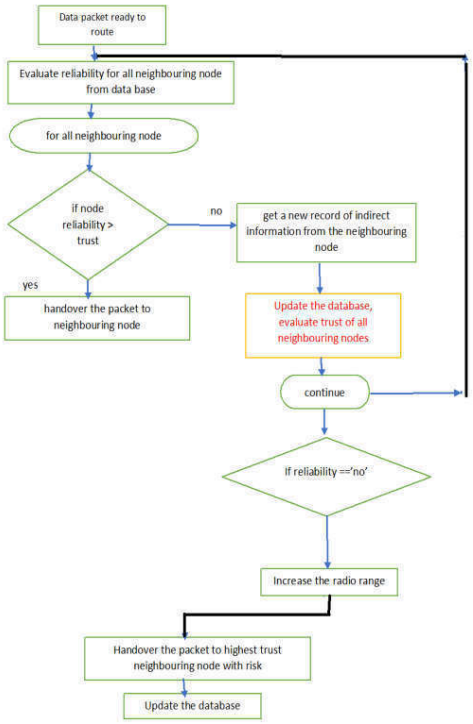


Fig-2 Reliability evaluation to route a packet

5 Implementation

Using our ERLEN model here we find the malicious node in WSN. We assume that a network having 13 nodes. Out of 13 nodes, we evaluated reliability on 5 node using instant node knowledge and oblique node knowledge. As we said earlier instant node will be categorized into two (based on the trust metrics) one is superior and another one is inferior. Oblique node is separated as high priority and normal node. Minimum threshold for ERELN T ≥0.15.after evaluating reliability if trust of a node less than 0.15 then the node will be detected as malicious.

In Table-2, ERELN detected node 2 and node 4 as a malicious node, because a trust of that two nodes is less than threshold level(0.15). The number of trust metric categories are taken is 10, out of which R1 to R5 are considered as superior trust metric and R6 to R10 are considered as inferior trust metrics of the node respectively.

Table-2 Detection of malicious node by ERELN

	Instant trust										Oblique trust									
Node	Superior					Inferior					High priority				Normal				Trust	Malicious node detection
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	D	E	F	G	H	I	J	K		
1	0.14	0.12	0.15	0.15	0.16	0.06	0.05	0.04	0.07	0.06	0.45	0.47	0.02	0.40	0.21	0.22	0.34	0.40	0.2394	
2	0.08	0.09	0.02	0.15	0.16	0.06	0.05	0.02	0.05	0.03	0.27	0.47	0.02	0.40	0.21	0.22	0.34	0.40	0.14500	Detected
3	0.14	0.12	0.15	0.15	0.16	0.06	0.05	0.04	0.07	0.06	0.45	0.47	0.42	0.40	0.21	0.22	0.34	0.40	0.254351	
4	0.14	0.12	0.11	0.10	0.12	0.06	0.05	0.04	0.05	0.06	0.22	0.23	0.02	0.40	0.21	0.22	0.34	0.40	0.1040941	Detected
5	0.12	0.14	0.16	0.15	0.15	0.05	0.04	0.06	0.07	0.06	0.45	0.47	0.42	0.40	0.21	0.22	0.34	0.40	0.157025	

Here we assumed that D, E, F, G are higher priority secondary node and H, I, J, K are
lower priority secondary node. Table-2 shows detection of malicious node by evaluating reliability through trust metric and commends given by the secondary node. We observed that the detection of malicious node using our technique ERLEN is highly depend on superior trust metrics of a primary node (instant trust) and weights given by the high priority secondary node (oblique trust).

Based on the ERLEN method on fixed WSN, malicious nodes are detected.

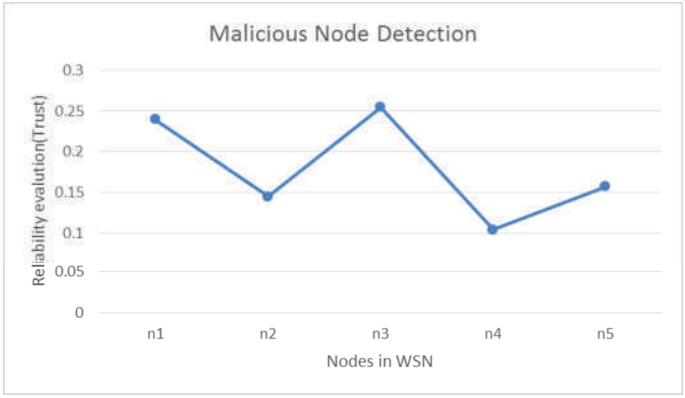


Fig-3. Reliable evaluation . This shows that node n2 and n4 are detected as malicious node. Other n1,n3,n5 are denoted as an reliable node.

6 Conclusions

The ERELN model proposed in this paper detects the malicious and/or selfish nodes in the network if any, and provides the reliability among the node and its neighbors based on the trust metrics evolved by former communications in the network and the commendations

provided by the neighboring nodes. The instant trust formed by trust metrics and oblique trust is formed by neighboring node recommendation. Reliability is evaluated based on instant trust and oblique trust. We have shown tabular form and graphically how ERLEN technique detects malicious node and find the reliable nodes. The proposed ERLEN technique will work for fixed WSN and it consumes time in densely populated wireless sensor network. In future we will extend our technique to work for dynamic WSN and reduces time consumption.

7 References

1 Lopez.J,Roman.R,Agudo.I and Fernandez-Gago,C(2010) Trust Management Systems for Wireless Sensor network:BestPractices.Computer Communication ,33,1086-1093.http://dx.doi.org/10.1016/j.comcom.2010.02.006.

2 R.Chinnaiyan, S.Somasundaram(2010) "Evaluating the Reliability of Component Based Software Systems " ,International Journal of Quality and Reliability Management , Vol. 27, No. 1., pp. 78-88.

3 R.Chinnaiyan, S.Somasundaram(2011), "An Experimental Study on Reliability Estimation of GNU Compiler Components - A Review", International Journal of Computer Applications,Vol.25, No.3, July 2011, pp.13-16.

4 R.Chinnaiyan, S.Somasundaram(2009) "Reliability of Object Oriented Software Systems using communication variables – a review " ,International Journal of Software Engineering, Vol.2, No.2,PP.87-96.

5 M.S.Nidhya, Dr.R.Chinnaiyan (2015) ," Reliability Oriented Protocols for Wireless Sensor Networks- A Review" , International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353, Vol.13 No.1, pp.491-494.

6 M.S.Nidhya, Dr.R.Chinnaiyan (2015) Improving the Reliability of data transfer against sinkhole in Wireless Sensor Networks – a Review, International Journal of Applied Engineering Research (IJAER), Vol.10, No.82, pp.122-125.

7 M.S.Nidhya Dr.R.Chinnaiyan (2017) Conniving energy and time delay factor to model reliability for wireless sensor networks, IEEE xplore Digital Library

8 Momani, M(2008)Bayesian Methods for modelling and Management of Trust in Wireless Sensor Network, Ph.D Thesis ,University of Technology, Sydney.

9 Sung-Jib Yim,Yoon-Hwa Choi(2012) Neighbour-Based Malicious Node Detection in Wireless Sensor Network,Wireless Sensor Networks, Vol.4,pp 219-225.

10 Trakadas.P.,Maniatis, S.Zahariadas,T.Leigou,H.C and Voliotis, S.(2009) A Novel flexible Trust management System for heterogeneous Sensor Networks. International Symposium on Autonomous Decentralised Systems,ISADS 2009, Athens,23-25 March 2009,369-374.

11 I.M Atakli, H.Hu, Y.Chen,"Malicious Node Detection in Wireless Sensor Networks Using Weighted Trust Evaluation," Proceedings of Spring Simulation Multi-conference,Ottawa,14-17 Apr 2008,pp.836-843.

12. M.S.Nidhya, Dr.R.Chinnaiyan (2017) "Shortest Reliable path for Wireless Sensor Network" International Journal of Pure and Applied Mathematics (IJPAM) Vol 117, No.20,pp.105-108.

13. M.S.Nidhya, Dr.R.Chinnaiyan (2018) " Reliability Evaluation on Wireless Sensor Network using EERN algorithm" Lecture Notes on Data Engineering and Communication Technologies vol15, 456649_1_En(78).

14. Dr.A.Kannagi, "K-Partitioned smallest distance Mining Tree for path Optimization in Wireless Sensor Network", Journal of Wireless Personal Communication - (2018). https://doi.org/10.1007/s11277-018-5996-7, Online ISSN1572-834X.

15. Dr.A.Kannagi, "Smart curiosity sink node prediction mining algorithm for path optimization in wireless sensor network", International Journal of Engineering and Technology(IJET) – Volume 7, No.2.21(2018) : Special Issue 21, Page No : 180-184.

16. A.Kannagi, Dr.Vidushi Sharma, Dr.K.Ganesan "Analysis of advanced data mining in search engine", International Journal of Emerging Innovations in Science and Technology(IJEIST), Vol.1, Issue – 1, March-2014, ISSN: 2348-4403.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Tackling Data Science for aliment forecasting in Health care

G NAGAPPA[1], G PRAVEEN KUMUR [2], R RAVI[3]

[1]*Assistant Professor, Department of CSE, Malla Reddy Institute of Engineering and Technology, Maisammaguda, Hyderabad, Telangana, India*

[2]*Assistant Professor, Department of CSE, Malla Reddy Institute of Engineering and Technology, Maisammaguda, Hyderabad, Telangana, India.*

[3]*Assistant Professor, Department of CSE, Malla Reddy Institute of Engineering and Technology, Maisammaguda, Hyderabad, Telangana, India*

ABSTRACT

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "; mined"; to discover hidden information for effective decision making. The analysis of healthcare parameters and the forecasting of the subsequent future health conditions are still in the informative stage. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. A probabilistic data collection mechanism is designed and the correlation analysis of those collected data is performed. A stochastic forecasting model is designed to foresee the future health condition of the most correlated patients based on their current health status.

Keywords: Data Science, Machine Learning, Disease Forecasting, Pruning Regression, Randam Forecasting Regression.

I INTRODUCTION

Over the last decade, there has been a rapid digitalization across the industries. Data Science analytics platform is the process to analyze the structured and unstructured data generated from healthcare management systems. A problematic information assortment mechanism is intended and therefore the correlation analysis of this collected knowledge is performed. A stochastic forecasting model is designed to foresee the future health condition of the most correlated patients based on their current health status. In healthcare communities, a large volume of multi structured patient data is generated from the clinical reports, doctor's notes and wearable body sensors.

The existing work focused on data type in the area of medical big data analytics. Compared with several typical forecasting algorithms, the forecasting accuracy of our proposed algorithm reaches 94.8% with a convergence speed, which is faster than that of the CNN-based modal disease risk forecasting algorithm. Imagine a system where doctors can quickly comb through millions of anonymzed patient records to find people with conditions and medical experiences. Through this massive, searchable database, doctors could determine how best to treat you, based on what has worked effectively for others with similar symptoms and characteristics.Stanford medicine is laying the groundwork for such a system, which will be able to quickly analyze information from large patient databases, medical literature, mobile monitoring and patient's real-life experiences with drugs, among other sources, to provide an evidence-based approach to medicine that's not been possible before.

Much of the highly valuable healthcare data is in unstructured or semi-structured form. Added to it, the complex, dynamic and heterogeneous characteristics of the data renders it difficult to extract useful information using traditional data

ISBN 978-81-939386-1-4

analytical tools & techniques. In fact, there is a finite human ability to process this data without effective decision support .This creates the need for integration of Big Data analytics into healthcare. The planned system is an example of how clinicians at Stanford medicine are tapping health data to provide targeted, predictive and personalized care, an approach known as precision health.

"Precision health is a way of translating data into information that can lead us to take care of our health in a way that we might not have done before," Minor said. "We are poised to have a whole new level of precision in maintaining health."

Precision health at Stanford Medicine has its roots in advances in both basic research and biomedical data science, which have given researchers the power to analyze vast quantities of information from a variety of sources: electronic medical records, genomic sequences, insurance and pharmaceutical records, wearable sensors and social and environmental data.

In shifting through this data, physicians and researchers can better predict individual risks for specific diseases, develop approaches to early detection and prevention, and arm clinicians with information to help them

make real-time decisions about the best way to care for patients.

Large-scale data analysis also is enabling researchers to develop more targeted and cost-effective methods for early diagnosis of disease before symptoms ever develop. Maximilian Diehn, MD, PhD, assistant professor of radiation oncology, and his colleagues did a study using a national library of DNA sequences of tumors to develop a technique sensitive enough to detect a single molecule of tumor DNA among thousands of healthy molecules. That means a blood sample could one day be enough to diagnose many types of solid cancers or to monitor the level of cancer in a patient's body.

In another recent study, School of Medicine researchers detected a pattern of gene activity that could lead to the development of a blood test to quickly and accurately detect sepsis, a deadly, whole-body inflammation syndrome. Sepsis is linked to 750,000 deaths each year in the United States and is the most common cause of hospital deaths. But it can be hard to pinpoint the condition, which can kill patients in a matter of hours. The researchers used publicly available patient data to identify specific immune-response genes associated with the condition, laying the groundwork for a simple diagnostic test for sepsis.

Which take advantage of Stanford's close ties to Silicon Valley, aim to use large collections of data to discern the factors that keep people healthy or send them along a disease. In collaboration with Google X and Duke University, Stanford researchers are conducting a pilot study, ultimately involving 10,000 healthy volunteers, to understand the biological markers of health, down to the molecular and cellular level.

The promising value of Big Data technology in healthcare has created an increasing interest of academic & industry

investigators. Nevertheless, there have been only a few literature reviews and the literature remains largely fragmented. The purpose of this research therefore is to gain a comprehensive understanding of current outlook on this technology.

"These partnerships with Silicon Valley innovators take advantage of new technologies, including mobile health devices, to do systematic studies of all the factors at play in maintaining a healthy state," Minor said. "That is the heart of precision health using all the resources at our disposal to not only diagnose and treat the sick, but to broaden our understanding so that we can better advise patients on how sustain their health throughout their lives.".

ISBN 978-81-939386-1-4

I RELATED WORK

Distributed Crowd filtering mechanism based on Heterogeneous Sensor Network

In the heterogeneous sensor network, due to the burst and random error, there are the transmission delay, convergence delay, the upstream and the downstream delay etc. because of the burst and random error. In this model, the sensor nodes are controlled independently and the signals are received or transmitted according to the change of environment. The coordinator is driven by the session requirement. Convergence nodes are often in periodic operation state or transient dormancy state. In a dormant period, the coordinator can be activated by a session or a server request. The bit error rate is random and burst. The bit error rate affects the data flow between the coordinator and the sensor. The bit error affects the data flow between the sensor and the sink node.

The transmission delay is the network delay between coordinator and sensor. the convergence delay is the network delay between the sensor and the sink node. The uplink communication delay between the sink node and the server is the uplink delay. Downlink delay is the delay of the controller

and the sink node. Assuming that network error will not result in the packet error. In a continuous network operating cycle T, a Pruning description of the network system is we use the coordinate signal matrix X (T) CO, the convergence signal matrix CG (T) x, and the control signal matrix CT (T) x computing network using the signal sequence x (t). We can obtain the system network delay according to the weight coefficient a1, a2, a3 and a4. Here, TD is transmission delay. CD is delay said protector. DLD is downlink delay. ULD is uplink delay. In order to maintain the consistency of network delay and the nature of signal attenuation, the sum of a1, a2, a3, and a4 must be 1. α is the lower weight of the working cycle of the network delay. β is the upper bound weight for the work cycle of network delay.

Planning foe the sustainability of community-based Health Program:

The changing economy and environment of human society, the acceleration of population ageing and chronic disease have become the top threat to human health. And the morbidity keeps rising. However, it's not easily accessible for people to get and take public health services because of the shortage in medical resources and facilities. Those medical resources and systems are focused in urban cities based on population and capital. Up to now, the medical and health systems in some countries (especially the developing countries) are mostly intended to cope with acute diseases.

Nowadays, the improvement of living standards pushes the rising of chronic disease. In US, 50 % of people suffer from one or multiple kinds of chronic diseases at different levels. 80 % of medical funds are used to treat chronic disease. In 2015, US Spends around $2.7 trillion on chronic disease treatment. This is accounts for 18 % of US GDP.

In Asia, the fiscal expenditure of Chinese Government on medical and health services has been increasing constantly, from more than eight hundred billion RMB in 2013 to more than one trillion RMB in 2014. The expensive medical expenditure brought enormous financial burden to society and local government.

In 2015, WHO produced a report on chronic diseases. It lists the four major chronic disease types, i.e., cardiovascular diseases, cancers, chronic respiratory diseases and diabetes mellitus. The report indicates that in 2012, most of non-communicable disease deaths was caused by these four diseases among under 70th age people. Cardiovascular diseases took the largest proportion of chronic death under the 70th age (37 %), followed by cancers (27 %), and chronic respiratory diseases (8 %). Diabetes took up 4 % death and other factors accounted for approximately 24 % of deaths.

The main causes of chronic diseases include three kinds of factors. They are unchangeable factors, changeable factors, and those factors hard to be changed. Age and heredity belong to the unchangeable factors and account for 20 % in the cause of chronic diseases. Living circumstances are critical for one's physical condition. And it is hard to be changed. In addition to environmental factor, worldwide social and economic trends, such as population ageing, urbanization, globalization, also bring impact on the cause of chronic diseases.

The population ageing is the direct reason of rising numbers of chronic diseases patients. And the urbanization makes the environmental pollution worse. For example, the heavy PM2.5 and haze have stimulated the increase of getting lung disease. Another aspect is the globalization. People are getting used to communicate with friends by mobile devices. The new technologies on social, mobile and networking are making more convenient urban life. However, these

phenomenon's give a rise of a variety of unhealthy lifestyle. For example, sitting in front of computer for too long time, lack of physical exercise to cause obesity problem, etc. These changeable factors are becoming the largest threat to modern people's health.

Using Analytics to Identify and Manage High-Risk and High-Cost Patients analytics:

it will be important to investigate analytic techniques that identify not only high risk people, but also those who are at particularly low risk. For instance, the standard approach may be to give all patients who are discharged from the hospital a follow-up appointment in two weeks. But it might make more sense to ensure that the highest-risk patients are seen within two days, while patients with very low risk might require follow-up care only as needed. Algorithms can help reallocate resources more effectively at both the high-risk and low-risk ends of the spectrum.

Programs to manage high-cost patients are expensive. They will be much more cost-effective if interventions can be precisely tailored to a patient's specific problems, which might be related to transportation, medication non-adherence, or family conflict.

Heart score to further risk stratify patients with low TIMI scores.

The ability to risk stratify patients presenting to the emergency department (ED) with potential acute coronary syndrome (ACS) is critical. The thrombolysis in myocardial infarction (TIMI) risk score can risk stratify ED patients with potential ACS but cannot identify patients safe for ED discharge. The symptom-based HEART score identifies very low-risk patients. Our hypothesis was that patients with a TIMI score of 0 or 1 may be stratified further with the HEART score to

identify a group of patients at less than 1% risk of 30-day cardiovascular events.

Conducted a secondary analysis of a prospective cohort study in a tertiary care hospital ED. Patients with potential ACS who were >30 years of age were included. Data collected included demographics, history, electrocardiogram, laboratories, and components of the TIMI and HEART scores. Follow-up was conducted by structured record review and phone. The main outcome was a composite of death, acute myocardial infarction,

At all levels of TIMI score, the HEART score was able to further sub stratify patients with respect to 30-day risk. A HEART score of 0 in a patient with a TIMI of 0 identified a group of patients at less than 1% risk for 30-day adverse events.

Data Science in health care: Using analytics to identify and manage high-risk and high-cost patients.

Health care system is rapidly adopting electronic health records, which will dramatically increase the

quantity of clinical data that are available electronically. Simultaneously, rapid progress has been made in clinical analytics techniques for analyzing large quantities of data and gleaning new insights from that analysis which is part of what is known as big data. As a result, there are unprecedented opportunities to use big data to reduce the costs of health care in the United States. We present six use cases that is, key examples where some of the clearest opportunities exist to reduce costs through the use of big data: high-cost patients, readmissions, triage, decompensating (when a patient's condition worsens), adverse events, and treatment optimization for diseases.

Discuss the types of insights that are likely to emerge from clinical analytics, the types of data needed to obtain such insights, and the

infrastructure analytics, algorithms, registries, assessment scores, monitoring devices, and so forth that organizations will need to perform the necessary analyses and to implement changes that will improve care while reducing costs. Our findings have policy implications for regulatory oversight, ways to address privacy concerns, and the support of research on analytics.

I Proposed Work:

A Data Science analytics platform is the best way to analyze the structured and unstructured data generated from healthcare management systems. A probabilistic data collection mechanism is designed and the correlation analysis of those collected data is performed. A stochastic forecasting model is designed to foresee the future health condition of the most correlated patients based on their current health status.

Machine learning algorithms are used for effective forecasting of chronic disease outbreak in disease frequent communities. The modified forecasting models over real-life hospital data collected. The analysis accuracy is reduced when the quality of medical data is incomplete. To overcome the difficulty of incomplete data, we use a Pruning regression model to reconstruct the missing data. To provide the best results we compare proposed algorithm with different forecasting algorithms. Analyze current and historical facts to make predictions about future unknown events. Predictive models exploit patterns found in historical and current data to identify risks and opportunities.

Models capture relationships among many factors to allow assessment of risk. Algorithms are identified to reduce the issues of existing system to improve the efficiency. Exploring of algorithms is going to be done to find the best results and improve the efficiency and performance. Incorrectness, complexity and time can be reduced by using

such algorithms. Systems that can reduce the issues relating to forecasting and handle data using suitable approach. Convolutional Neural Network based multimodal disease risk forecasting algorithm is used for both structured and unstructured data.

There is a large number of missing data due to human error. Prunning regression model isused to fill the missing data. Identify uncertain or incomplete medical data and

I IMPLEMENTATION

Implementation of modules deals with forecasting and analysis. Hospital dataset used in the process. It performs statistical observation on the dataset. Finally forecasting is performed on the training dataset.

Forecasting based healthcare applications, including health risk assessment. Forecasting using disease risk models usually involves machine learning algorithm (e.g., Pruning regression and random forest.)

ISBN 978-81-939386-1-4

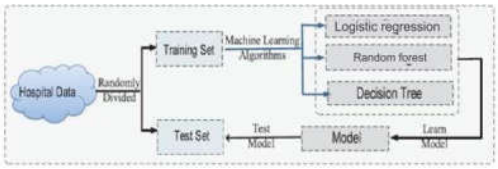


Figure:Machine learning algorithms used in disease prediction.

It explains the procedural of analysis the data, by classifying the data to testing data and training data. First we extract the data, then we apply Prunning regression model on the data to acquire the important and find the combination of the columns of the data. Later on the logistic regression model is applied on the data so that the forecasting can be obtained and the model is again applied with help of random forest algorithm which

then modify or delete them to improve the data quality. A statistical approach is framed from the data. Forecasting using disease risk models usually involves machine learning algorithm (e.g., logistic regression and random forest.) Usage of Machine learning algorithms for effective forecasting of various diseases. Use of Pruning regression model to overcome the difficulty of incomplete data.

results us with comparison model and accuracy model for the prediction.

Analytical Console: Data is loaded to the console and organized; cleaning of the data is simplified in necessary. The hospital dataset used in this data contains real-life hospital data, and the data are stored in the data centre to protect the patient's privacy and security.

By using statistics on the number of patients, the sex ratio of patients and the major disease in this region every year from the structured and unstructured text data, to realize the machine learning and deep learning algorithms and run it in a parallel fashion by the use of data centre.

Heart disease forecasting: Machine learning algorithms to predict heart disease (Angiographic disease status). The main objective is predicting the heart disease of a patient using machine learning algorithms.



"e1071" - Used for the Pruning regression model for assuming the required values from the dataset.

"pROC" - Assuming the statistical behavior of the data in terms of compering the gender and age wise distribution of the data.

"caret" - Holds the methods for the usage of the algorithms such as Pruning Regression, logistic Regression.

"Random Forest" - Method of random forest are applied with this library, implementation on the data.

Statistical Observation: Based on the data sets, a statistical approach is framed from the data. There is a large number of missing data due to human error. Thus, we need to fill the structured data. Before data imputation, First identify uncertain or incomplete medical data and then modify or delete them to improve the data quality. Using Pruning regression model to find the missing data. We use data integration for data pre-processing. To explain the variables in terms of the latent variables.

Algorithm for Pruning regression model:

Step 1: Load the data to console. Step 2:

Check for any missing values

If any missing values: replace

ISBN 978-81-939386-1-4

them with 0 Then step 1

else Continue

Step 3: Sort the combination attributes. Step 4:

Correlation over the attributes. Step 5: Test the

data.

Training Dataset: From all the training dataset, we apply the random forest algorithm in finding all the possible decision trees and predict the results. Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. The decision tree algorithm is also implemented.

Creating more number of decision trees. The calculation of nodes selection will be same for the same dataset. The same algorithm can use for both regression and classification problem.

Algorithms for Pruning regression and random forest:

Step 1: Train the data on minimum dataset. Step

2: Check the results.

Step 3: Apply on the complete dataset.

Step 4: Generate the results from the model.

Form the above results generate the reports in terms of behavior of the data. Missing values problem can be resolved using machine learning algorithms.

II EXPERIMENT AND RESULTS

The implementation is executed with the sample data available from a local website. The data is first categorized and then the core parameters are been identified from the dataset. The statistical analysis of the data is executed. Once the parameters are identifiable later we have implemented with logistic regression model for the forecasting of the heart diseases and compared our model with the Random forest model for the accuracy calculation. In this process the data is been categorized into testing data and training data and found out the results to be accurate and found that the logistic model provided better results compared with random forest.

Data category	Item	Description
Structured data	Demographics of the patient	Patient's gender, age, height, weight, etc.
	Living habits	Whether the patient smokes, has a genetic history, etc.
	Examination items and results	Includes 682 items, such as blood, etc.
	Diseases	Patient's disease, such as cerebral infarction, etc.
Unstructured text data	Patient's readme illness	Patient's readme illness and medical history
	Doctor's records	Doctor's interrogation records

HOSPITAL DATA

The hospital dataset used in this study contains real-life hospital data, and the data are stored in the data center. To protect the patient's privacy and security, we created a security access mechanism. The data provided by the hospital include EHR, medical image data and gene data. We use a three year data set from 2013 to 2015. Our data focus on inpatient department data which included 31919 hospitalized patients with 20320848 records in total. The inpatient department data is mainly composed of structured and unstructured text data.

The structured data includes laboratory data and the patient's basic information such as the patient's age, gender and life habits, etc. While the unstructured text data includes the patient's narration of his/her illness, the doctor's interrogation records and diagnosis, etc. As shown in Table 1, the real-life hospital data collected from central China are classified into two categories, i.e., structured data and unstructured text data. In order to give out the main disease which affect this region, we have made a statistics on the number of patients, the sex ratio of patients and the major disease in this region every year from the structured and unstructured text data, the statistical results are as shown in Table II. From Table 2, we can obtain that the proportion of male and female patients hospitalized each year have little difference and more patients admitted to the hospital in 2014. Moreover, the hospitalization resulted by chronic diseases has always been occupying a large proportion in this area through the statistics of the data. For example, the number of patients hospitalized

Table 1:Types of hospital data.

Statistics	2013	2014	2015
Number of inpatients	7265	24756	10552
Males	42.88%	50.36%	57.60%
Females	57.12%	49.64%	42.40%
Proportion of patients with cerebral infarction	1.47%	1.01%	1.66%
Proportion of hypertensive patients	1.06%	1.04%	1.98%
Proportion of diabetics	1.17%	0.99%	1.99%

Table 2: Initial statistics from hospital data.



Figure: Representing the hospital dataset.

Hospital dataset is represented with variables. Data is loaded and viewed in the console. Dataset is represented with variable names. Short description of the variable names is generated.

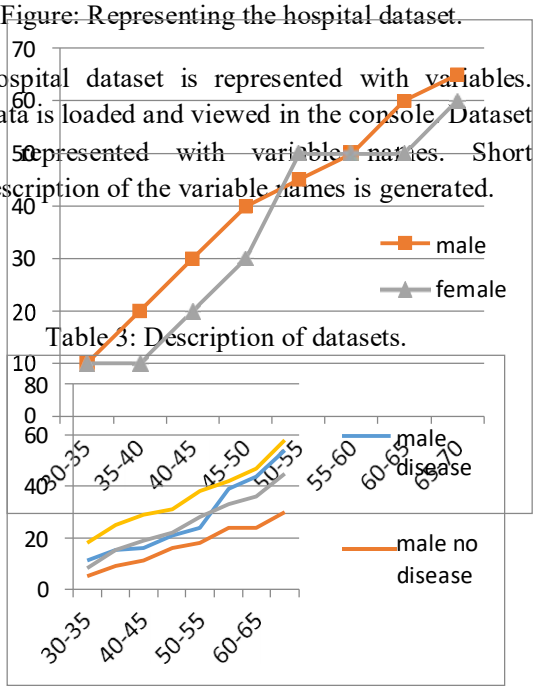


Table 3: Description of datasets.

It represents about the variable name and



Variable name	Short description	Variable name	Short description
age	Age of patient	thalach	maximum heart rate achieved
sex	Sex, 1 for male	exang	exercise induced angina (1 yes)
cp	chest pain	oldpeak	ST depression induc. ex.
trestbps	resting blood pressure	slope	slope of peak exercise ST
chol	serum cholesterol	ca	number of major vessel
fbs	fasting blood sugar larger 120mg/dl (1 true)	thal	no explanation provided, but probably thalassemia (3 normal; 6 fixed defect; 7 reversable defect)
restecg	resting electroc. result (1 anomality)	num	diagnosis of heart disease (angiographic disease status)

description of the attributes. which are present in the dataset

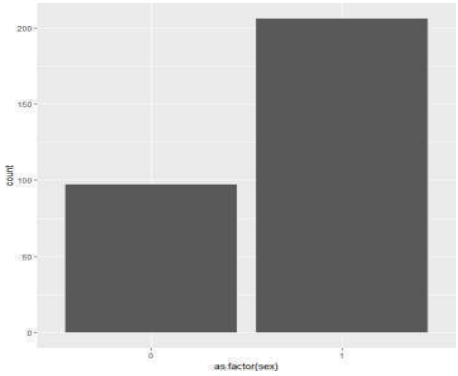


Figure: Representing thebarplot using dataset

The figure describes most effected gender for the given dataset. As shown in the above figure mostly male is effecting the disease.

ting the forecasting accuracy

Representing thaccuracy figure for training and testing dataset As shown in the above figure, it describes about the forecasting accuracy for training and testing datasets

	Logistic Regression	Random Forest
Accuracy 70	0.86516	0.80898
Accuracy 60	0.86440	0.84747
Accuracy 50	0.72729	0.80405

Figure: Representing most effected gender

Most effected gender for the combination of attributes which are represented in the dataset.

Figure: Representing most affected age group.

It represents which age group of the people is most effected in the given dataset.

It represents the forecasting accuracy for the dataset with different algorithms which are performed at various intervals of the dataset with training and testing data.

III CONCLUSION

Analysing the disease factors based on the health care data a random forest model based multimodal disease risk forecasting algorithm using structured data from hospital. Much of the highly valuable healthcare data is in unstructured or semi-structured form. Added to it, to the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Analysing the disease factors based on the health care data. Missing values problem can be resolved using machine learning algorithms. Disease forecasting can be done based on type of the disease and type of disease, region and risk level of the patient's disease by the availability of the data.

IV REFERENCES

] G. Fortino et al., "A Framework for Collaborative Computing and Multi-Sensor Data Fusion in Body Sensor Networks", *Information Fusion*, vol. 22, pp. 50-70,
2015.

[2] van Kuiken, *The `Big Data' Revolution in Healthcare: Accelerating Value and Innovation*. USA:Centres for US Health System Reform Business Technology Office, 2016.

[3] P. B. Jensen, L. J. Jensen, and S. Brunak, ``Mining electronic health records: Towards better research applications and clinical care," *NatureRev. Genet.*, vol. 13, no. 6, pp. 395_405, 2012.

[4] M. Chen, Y. Ma, J. Song, C. Lai, and B. Hu, ``Smart clothing: Connecting human with clouds and big data for sustainable health monitoring," *ACM/Springer Mobile Netw. Appl.*, vol. 21, no. 5, pp. 825-845,2016.

[5] M. Chen, S. Mao, and Y. Liu, ``Big data: A survey," *Mobile Netw.Appl.*, vol. 19, no. 2, pp. 171_209, Apr. 2014.

[6] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, ``A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular

ISBN 978-81-939386-1-4

Survey on women safety based on IOT

Thamaraiselvi K[1] , Assistant Professor, Malla reddy College of Engineering, Telangana.

Rinesh S[2] ,Assistant Professor,Saveetha University,Tamilnadu

Kalpana A[3], Assistant Professor, Sri Saradha College for Women, Salem, Tamilnadu

Abstract:

Sexual abuse is a major problem in India. Now a day's crime against female children has been increased to great extent, security and safeties are major issues of female. According to survey has found that more than 53% of female children in India subjected to sexual abuse and those in the age group of 5-12 years reported higher level of abuse. In this paper we survey about various type of women safety wearable band based on IOT application.

Keywords: ***IOT, Safety wearable band, sexual abuse.***

Introduction

Day to day woman has been facing lot of security challenges in the world. In last few years victims against women in India were increased.[1] In India the survey revealed that one in every two children is a victim of child sexual abuse.[2],Even though female child were not allowed to move alone anywhere like streets, tuition, travelling by bus etc. current global scenario, the prime question in every girl's mind, considering the ever rising increase of issues on women harassment in recent past is mostly about her safety and security. 848 Indian Women Are Harassed, Raped, Killed Every Day!!" That's a way beyond HUGE number. In this paper discussed with various smart and efficient intelligent security system for women, the system can be built that can detect the location and health condition of person that will enable us to take action accordingly based on electronic gadgets like GPS, body temperature sensor , pulse rate sensor ,motion sensor and IOT. [3]

Literature survey

(Roryyev A et al, 2011.)[4] Present a fuzzy-based indoor tracking sys- tem for localizing object in an environment characterized by Wireless Sensor Network (WSN). WSN sensor node detects if there is any target node and receives a response. Upon the receipt of the reply, the sensor node calculates the times for receiving signal and the return trip to the target as well as the strength of the signal received. Fuzzy logic model was applied for obtaining the x-y coordinates of the different nodes. The major limitation of the system is its failure to explicitly define the identity of the object.

(Harshitha et al)[2016][3] proposed smart band connected to smart phone

through wifi &in this system provide security to women while facing social issues.

(Dongare Uma et al, 2015.) [5]Discussed the voice recognizing module will recognize the user and activates the app even if the mobile keypad is locked. The GPS in mobile will read the latitude and longitude to trace the exact location of the user and sends the message alert including the location share to registered contact numbers. The audio recording module records the conversation for evidences. Also user can select contact number to make a call through voice based contact list.

(Ravi Sekhar et al, 2015)[6] even a single click in the mobile app it first make a call to the first registered contact number and sends the alert message to all the registered contact numbers along with sharing of location continuously for every five minutes until press down stop button.

on the application.

(Shreyas et al, 2016.)[7] When the user press the control button present in smart watch it automatically recognize and sends the voice command to the nearest police with the current location. It also contains a shock mechanism to produce electric shock in emergency situations to detect the attacker.

(Helan et al , 2017.)[8] in this paper discussed the sensor in the watch gets activated when the targeted heart rate for the targeted time period is achieved and produces alarm sound to make the nearby people attention. It automatically calls the registered contact list and sends the user's location through message.

(N.Vijaykumar et al , 2017)[9] proposed a smart intelligent device which automatically senses information and helps women in "Every single step of life". The device is the integration of multiple devices comprises of a wearable smart-band and a secret webcam connected via Bluetooth which continuously tracks the information and communicates with smart phones that has access to the internet. The application is programmed and embedded in such a way that tracks information of the women such

as call log, messages, pulse measurement, movement, , heart beat rate ,blood oxygen levels and records continuously in the internet. When SOS present in smart band is pressed continuously it automatically generates signals to the predefined smart phones and nearest police station along with location coordinates and the secret webcam in the locket captures the culprit photo which is directly uploaded to the server.

(Jain et al,2017.)[10] Proposed different sensors like heartbeat sensor, accelerometer sensor and temperature sensor for detecting temperature of the body, heartbeat and sudden change in motion of

user. The GPS used to detect location of the device & The GSM in the model is used to send alert message to guardians, relatives and which is nearest police station.

(Thavil et al,2017.)[11] developed one android application which is useful as women security purpose. Generally this service can be start by clicking the ACTIVATION SERVICE BUTTON. By pressing VOLUME KEY emergency message will be send to the registered contacts. "Design And Development of Women Self Defence Smart Watch Prototype (More et al ,2017)[12]Developed android application for women safety wearable devices that will transmit data for comparing with the original dataset, if the values are irregular based on body temperature, pulse rate are identified then alert message will be sent to family member registered contact list and location based nearby police station.

(Akinyokun O et al, 2018.)[13] presented the integration of smart technologies and Internet of Things to track criminals and promote safety of lives and properties. In this IoT basesystem for achieving effective cost and time saving combat of kidnapping and human tracking. The Smart Engine Server will perform logical reasoning of the system such as face recognition, motion detection, location tracking and activities recognition; Digital Situation Room will be used to monitor events in real-time and on-demand modes based on IPv6 enabled communications.

(Pramod et al, 2018.)[14] In this paper discussed with wearable IOT device for the security and shielding of female children. This is accomplished by the examination of physiological signs in concurrence with body gestures. In the signs are analyzed along with body temperature is measured by galvanic skin resistance. This work deals with body temperature and skin resistance. Movements of the persons based on position and location is recorded. The device makes an analysis of skin resistance and body temperature to analyze the situation of the person.

Conclusion

In current environment the prime question in every female mind, considering the ever rising increase of issues on women harassment in recent, is mostly about her safety and security. The above literature proposed various aspects of Women Safety and security based on IOT based wearable band. But Safety planning is about brainstorming ways to stay safe that may also help to reduce the risk of future harm. Be a parent explain private parts of body , monitoring child behavior, don't allow mobile phone for using, good and bad touch of the body and self defense action course(like karate and kung fu) to allow for both female and male children.

Reference

[1] http:www.womenhealth.gov/realtionships
-and-safety/sexual-assault-and-rape. Kavya et al, " IOT Based Women's Safety Gadget",IJIRSET, pp.286- 289,Vol 7,Special Issue 1, March 2018.

[2] Haritha et al"Smart Security Solution For Women Using IOT ", Project Reference No: 40S_BE_2029,2015.

[3]Roryyev A, Hasbullah H, Subhan F , "Indoor Child Tracking in Wireless Sensor Network using Fuzzy Logic Technique Research". Journal of Information Technology,pp.81-92,Vol 3(2),2011.

[4] Dongare Uma, V yavahare Vishakha and Raut Ravina, "An Android Application for Women Safety B ased on Voice Recognition", International Journal of Computer Science and Mobile Computing (IJ CSMC), pp. 216-220,Vol.4 Issue.3,March- 2015 .

[5] Ravi Sekhar Yarrabothu and Bramarambika Thota, "Bhaya: An An Droid App For The Safety Of Women", 978-1- 4673-6540-6/15©2015 IEEE

[6] Shreyas R.S, Varun.B.C, Shiva Kumar.H.K, Punith Kumar B.E, Kalpavi.C.Y, "Design And Development Of Women Self Defence Smart Watch Prototype",IJARECE,pp.1179-1185, Volume 5, Issue 4, April 2016

[7] A. Helen , M. Fathima Fathila, R. Ri jwana and Kalaiselvi .V.K.G, "A Smart Watch For Women Security Based On IOT Concept Watch Me " , 978-1-5090- 6221-8/17/@2017 IEEE.

[8] N.Vijaykumar et al, "Efficient Tracking For Women Safety And Security Using IOT ",International Journal of Advanced Research in Computer Science, pp.328-330 8(9), Nov–Dec, 2017.

[9] Prof. R.A.Jain et al, "Women's safety using IOT", International Research Journal of Engineering and Technology (IRJET), pp.2336-2338,Volume: 04 Issue: 05 ,May -2017.

[10] J.K.Thavil1, V.P.Durdhawale, P.S.Elake, "Study on Smart Security Technology for Women based on IOT", International Research Journal of Engineering and Technology (IRJET) pp.1099-1101, Volume: 04 Issue: 02 | Feb -2017

[11] S. A. More , R. D. Borate, S. T. Dardige, S. S. Salekar , Prof. D. S. Gogawale , "Smart Band for Women Security Based on Internet of Things (IOT)" ,International journal of advanced engineering , pp 840-845,Vol 6,issue 11,2017.

Akinyokun O et al"Design of a framework for combating human tracking and kidnapping using smart objects and Internet-of-things", Artificial Intelligence Research, pp.45-52,Vol. 7, No. 1,2018.

[12] M. Pramod, Ch V. Uday Bhaskar and K. Shikha, "IOT Wearable Device For The Safety And Security Of Women And Girl Child",International Journal of Mechanical Engineering and Technology (IJMET), pp. 83–88,Volume 9, Issue 1, January 2018.

[13] Thamaraiselvi K and Balamurugan P.S., "Performance Analysis of Homogeneous and Heterogeneous System Using Bit Torrent in Compeer Networks",(IJCSIT)International Journal of Computer Science and Information Technologies, pp.3821-3824,Vol.3(2),2012.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A new framework for video data retrieval using hierarchical clustering technique

D.Saravanan
Faculty of Operations & IT
ICFAI Business School (IBS), Hyderabad,
The ICFAI Foundation for Higher Education (IFHE)
(Deemed to be university u/s 3 of the UGC Act 1956)
Hyderabad-India.

ABSTRACT

The amount of information produced every year is rapidly growing due to many factor among all media, video is a particular media embedding visual, motion, audio and textual information. Given this huge amount of information we need general framework for video data mining to be applied to the raw videos (surveillance videos, news reading, Person reading books in library etc.).We introduce new techniques which are essential to process the video files. The first step of our frame work for mining raw video data in grouping input frames to a set of basic units which are relevant to the structure of the video. The second step is charactering the unit to cluster into similar groups, to detect interesting patterns. To do this we extract some features (object, colors etc.)From the unit. A histogram based color descriptors also introduced to reliably capture and represent the color properties of multiple images. The preliminary experimental studies indicate that the proposed framework is promising.

KEYWORDS: Data Mining, Video Segmentations, Video Data Mining, Clustering, Histogram, Video Data Clustering.

1. INTRODUCTION:

Data mining, which is defined as the process f extracting previously unknown knowledge, and detecting interesting patterns from massive set of data, has been a very active reseach.As result, several commercial product and research prototypes are even available nowadays. Multimedia data mining has been performed fro different types of Multimedia data: image, audio and video [1]. An example of image data mining is CONQUEST [2] system that combines satellite data with geophysical data to discover patterns in global climate change. The SKICAT systems[3] intergrades techniques for image processing and data classification in order to identify 'sky objects' captured in a very large satellite picture set. The MULTIMEDIAMINER [4] project has constructed many image understanding, indexing and mining techniques in digits media. Digital multimedia differs from previous forms of combined media in that bits representing, text, images, audios, and videos can be treated as data by computer programs. One facet of these diverse data in terms of underlying models and formats is that they are synchronized and integrated hence, ca be treated as integrated data records. The collection o such integral data records constitute multimedia data set. The challenge of extracting meaningful patterns form such data sets has lead to research and development in the area of multimedia data mining. This is a challenging field due to the non-structured nature of multimedia data. Such ubiquitous data is required in many application such an s financial, medical, advertising and command, control, communication and intelligence. In Video Data mining, data bases are widespread and multimedia data sets are extremely large, there are tools fro managing and searching within such collections, but t the need for tolls to extract hidden and useful knowledge embedded with in data is becoming critical for much decision – making applications.

2. VIDEO DATA MINING:

Here we define video data mining as finding correlations and patterns previously unknown, the current status of video data mining remains mainly at the pre-processing stage, in which the preliminary issues such as video clustering, and video classification are being examined and studied for the actual mining.

2.1 Requirements of a Video Mining System:

The following requirements for a video mining system [5].

1. It should be unsupervised.
2. It should not have any assumption about the data.
3. It should uncover interesting events.

Note that requirements 2 and 3 are somewhat contradictory, since the notion of 'interesting' is subjective, and highly dependent on knowledge of the content. Our aim is to find out how few assumptions can make about the content without detecting events that are too general to be meaningful. This would help us understand the content, and help set up framework of systematic use of domain knowledge

2.2 Video Classification:

Classification is a way to categorize or sassing class labels to a pattern set under the supervision. Decision boundaries are generated to discriminate between patterns belonging to different classes. The data set is initially partitioned into segments and the classifier is trained on the former. A framework to enable semantic video classification and indexing in a specific video domain was proposed. A method for classification of different kinds of videos that uses the output of a concise video summarization technique that forms a list of key frames was present.

2.3 Video Clustering

Clustering is useful technique for the discovery of some knowledge from data set. It maps a data item into one of several clusters, where clusters are natural grouping for data items based on similarity metrics or probab8ility density models. Clustering consists of partitioning data into homogeneous granules or groups, based on some objective function that maximizes the inter-cluster distance. Video clustering has some differences with convential clustering algorithms, As mentioned earlier, due to the unstructured nature do video data, preprocessing of video data by using image processing g or computer vision techniques is required to get structured format features, Another difference in video clustering is that the time factor should be considered while the video data is processed.

Traditional clustering algorithms can be categorized into two main types: Partitioned and hierarchical clustering [6].Hierarchical algorithms determine all cluster at once. Hierarchical algorithms can be agglomerative ("bottom up") or divisive ("top down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Two-way clustering, co-clustering or bi clustering are clustering methods where not only the objects are clustered but also the features of the objects, i.e. if the data is represented in a data matrix, the rows and columns are clustered simultaneously.

2.4 Pattern Finding:

A general framework for video data mining was proposed to address the issue of how to extract previously unknown knowledge and detect interesting patterns. In the work, how-to segment the incoming raw video stream into meaningful pieces, and how to extract and represent some feature (i.e. motion, object difference, text) for characterizing the segmented pieces. Then, the motion in a video sequence is expressed as an accumulation o f quantized pixel differences among all frames in the video segment's video data base management framework and its strategies for vide content structure and events its strategies for video content

structure and events mining were introduced. Fig 1 shows the proposed framework.

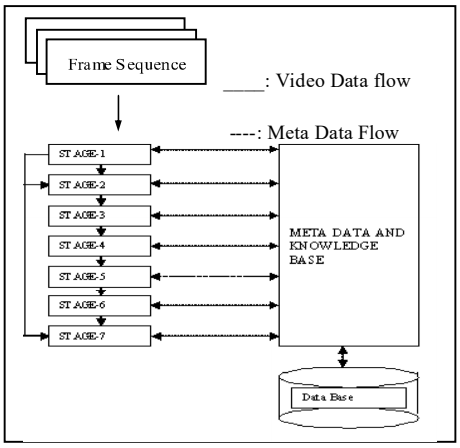


Fig 1: Proposed Framework for Video Data Mining

STAGE 1: The first stage of the frame work for mining raw video data is grouping input frames to a set of basic units which re relevant to the structure of the video. This is one of the most important takes since it is the first step to construct the building blocks of the video data base.

STAGE 2: The SECOND STAGE extracts the features such as motion, object colors etc to characterize these segments. For our frame work we consider three features (motion, object, and colors) extracted from each segments. To extract motion we make calculations using quantized pixel difference among all frames in segment.

STAGE 3: In stage THREE unwanted segments are removed by using clustering technique in clustering we employ multilevel hierarchical clustering approach to group segments with similar categories in the top level and similar motion in the bottom level.

STAGE 4: The NEXT STAGE is actual mining of raw video sequence using CLUSTERING technique.

STAGE 5: In this stage the actual mining of raw video sequences processed in the above stages, and video data compression for storage of these raw videos.

STAGE 6: In the stage 6 and 7 we going to create a indexing for the video file. The meta data & Knowledge Base in the fig a module to store the results from each stage and provide the necessary information to the stage.

3. IMAGE MINING

There are two major issues that will affect the image data mining process. One is the notion of similarity matching and the other is the generality of the application area. For a specific application area[7], associated domain knowledge can be used to improve the data mining task. Since data mining relies on the underlying querying capability of the CBIR system, which is based on similarity matching, user interaction will be necessary to refine the data mining process. With image mining we will consider the four broad problem areas associated with data mining: Finding associations, Classification, Sequential pattern and Time series pattern. With all these , the essential component in image mining is identifying similar objects in different images

3.1 Image Mining Algorithm Steps:

The algorithms needed to perform the mining of associations within the context of image. The four major image mining steps are as fallows:

1. Feature Extracting: Segment images into regions identifiable by region descriptors (blobs) ideally one blob represents one object.
2. Object identification and record creation: Compare objects in one image to objects in every other image. Label each object with an id. We call this step the preprocessing algorithm.
3. Create auxiliary images: Generate image with identified objects to interpret the association rules.
4. Apply data mining algorithm to produce object.

4.HISTOGRAM SEARCH

Histogram search algorithms [8],[9] characterize an image by its color distribution or histogram. A histogram is nothing but a graph that represents all the colors and the level of their occurrence in an

ISBN 978-81-939386-1-4

image irrespective of the type of the image. Few basic properties about an image can be obtained from using a Histogram. It can be used to set a threshold for screening the images. The shape and the concentration of the colors in the histogram will be the same for similar objects even though they are of different colors..

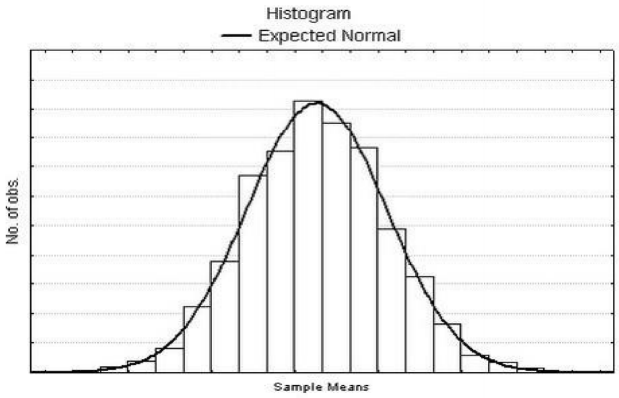


Fig 2. By using Histogram Data Analysis

4. EXPERIMENTAL RESULTS

Different experiments were performed, one for assessing computational performance, and one assessing robustness with respect to validate the methods we have described, we implemented the components of the video frame based retrieval system and tested with a general purpose image database including about 100 videos[10]. The table given below shows the time taken for Splitting Number of frames from the image database

Number Of Frames Spitted	Search Time
100	2 Seconds
1000	5 Seconds
10000	14 Seconds

Tab 1: Experimental Result of Spitted Frames

. Number Of Frames	Videos
13116	Cartoon
27956	Tennis
71379	News
13254	Movie Song

Tab 2: Avg Search Time for Particular Frame

Mathematical operations are finally conducted in the following manner:

First a codebook of frame – tregions is created on a set of training videos.

The co-occurrence matrix is constructed:

Let A of size M by N be the co-occurrence matrix of M centroids (defining a codebook) into N- key frames (representing the video database).its value at cell (i, j) corresponding to the n8mber of times the region I appears in the frame j [11].

Next, it is analyzed through LSA:

The SVD decomposition gives $A = USV^i$ where $\mathbf{UU^i = VV^i = I, L = min\ (M,N)S \approx diag\ (\sigma_1,\ldots\ldots\ \sigma_L\ ),\ \sigma_1 \geq \sigma_2 \geq\ldots\geq\sigma_L}$ Then A is approximated by truncating U and V matrices to keep k factors in S corresponding to the highest singular values.

$$\mathbf{\hat{A} = U_k S_k V_k^t \text{ with } S_k = diag(\sigma_1,\ldots\ldots\ \sigma_k)}$$

Finally, indexing of a context of a noted c (j) and a new context q is realized as follows:

$P_{C\ (j)}$ = row j of VS

$$\mathbf{U_p = q^t U_k}$$

And to retrieve the context q in a database containing indexed contexts $J_{PL,}$ the cosine measure m_c is used to compare elements.

$m_c(P_j,q) = P_q.P_j \ / \ \| P_q \|. \| P_j \|$ the most similar elements to the query are those with the highest value of m_c.

The number of singular values kept for the projection drives the LSA performance. On one hand too many factors are kept, the noise will remain and the detection of synonyms and polysemy of visual terms will fail. On the other hand if too few factors are kept, important information will be lost degrading performances. Unfortunately no solution has yet been found and only experiments allow finding the appropriate factor number.

In the particular situation of video content, many features can be extracted



Fig 3 the model Input Frame

6. CONCLUSION:

In this paper, we propose a general framework for video data mining to perform the fundamental tasks which are temporal segmentation of video sequences, feature extraction, and clustering of segments. Although the experimental data showing that the proposed framework is performing the fundamental tasks effectively and efficiently.

7. REFERENCES:

[1] Jung Hwan Oh, Babitha Bandi. Multimedia data mining framework for raw video sequences. In Proc. Of International Wrokshop on Multimedia Data Mining (with ACM SIGKDD 2002) , Pages 01-10, July 2002.

[2]P.Stolorz, H.Nakamura, E.Mersrobian, R.Muntz, E.Shek, J Santos, J Yi, K Ng, S.Chien, C.Mechoso, and J.Farrara. Fast spatial-temporal data mining of large geophysical data sets. In Proc. Of Int'l Conf. on KDD ,pages 300-305, 1995.

[3]D.Saravanan, "Video data image retrieval using – BRICH", World journal of Engineering, Vol.14, Issuu 4, Pages 318-323, Aug 2017.

[4]D.Saravanan," Image frame mining Using indexing technique" Data Engineering and Intelligent Computing , SPRINGER Book series, Chapter 12 , Pages 127-137, ISBN:978-981-10-3223-3,July 2017.

[3]U .Fayyad, S.Djorgovski, and N.Weir, Automating the analysis and cataloging of sky serverys. Advances in Knowledge Discovery with Data Mining, Pages 471-493, 1993.

[4]Z .N Li, O.R.Zaiane, and Z.Tauber. Illumination invariance and object model in Content-based image and video retrieval. Journal of Visual Communication and Image Representation, 1998.

[5] Ajay Divakaran, Dadir A.Peker, Shih-Fu-Change, Regunathan Radhakrishanan, Lexing Xie. Video Mining: Pattern Discovery versus pattern Recognition. In Proc. Of Int'l Conference on ICICS ,2006.

[6] Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma, " A survey of Content –based image retrieval with high-level semantics", The journal of Pattern Recognition Society, Vol 40, ,pp 262-282,2007.

[7] R. Benmokhtar and B. Huet,"Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content," International MultiMedia Modeling Conference, vol. 4351, pp. 196–205, Singapore, 2007.

[8]D.Saravanan" Effective Video Data Retrieval Using Image Key frame selection", Advances in Intelligent Systems and computing , Pages 145-155,jan-2017.

[9]D.Saravanan, "Clustering of video in formations using BRICH Methodology" Pak journal of biotechnology, Vol 14(Special Issue-II-2017), Pages 377-380, Dec- 2017

[10] Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T.(2004). Probabilistic Author-Topic Models for Information Discovery. The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington. 113–141, 2000.

[9D.Saravanan, Dr.S.Srininvasan Video Data mining pattern discovered using statistical modeling approach. In the proc. Of Int'l Confrence on Scinece and Applications ICISA 2010.,pp 353-355.

[10]Nevenka Dimitrova, Hong-Jiang Zhang, Behzad Shahraray, Ibrahim Sezan, Thomas Huang, and Avideh Zakhor, "Applications of video-content analysis and retrieval," IEEE Multimedia Magazine, vol. 9, no. 3, pp. 42 – 55, July 2002.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Video data retrieval using image color histogram technique

D.Saravanan
Faculty of Operations & IT
ICFAI Business School (IBS), Hyderabad,
The ICFAI Foundation for Higher Education (IFHE)
(Deemed to be university u/s 3 of the UGC Act 1956)
Hyderabad-India.

Abstract– In this paper, a subspace-based multimedia data-mining framework is proposed for video semantic analysis; specifically Current content management systems support retrieval using low-level features, such as motion, color, and texture. The proposed frameworks achieves full automation via a knowledge-based video indexing and retrieve an appropriate result, and replace a presented object with the retrieval result in real time. Along with this indexing mechanism a histogram-based color descriptors also introduced to reliably capture and represent the color properties of multiple images. Including of this a classification approach is also carried out by the classified associations and by assigning, each of them with a class label, and uses their appearances in the video to construct video indices. Our experimental results demonstrate the performance of the proposed approach.

Index Terms– video indices, video indexing, histogram

1 INTRODUCTION

The usage of the images has been increasing day by day as they are not only being used for expressing and explanation but also for identification and analysing. They are being used in every field today. Large amounts of image data is being stored in the databases everyday. All the data can only be used efficiently only when the correct data is retrieved successfully. The existing successful systems are text based and are not effective at retrieving the images. For efficient retrieval the image properties are used to retrieve the appropriate images which do not depend on image descriptions. The present day there are only a couple of systems that implement the retrieval mechanism based on content. The proposed system identifies the objects in the images based on the colour distribution and performs the search. A colour histogram is generated and image segmentation is done to obtain the suitable images. Every image is made up of tiny bins called pixels and each pixel is made up of the Red, Green and Blue colours. The system utilizes these values in retrieving the suitable solutions.

1.1 Related Work in Video Indexingz

The evolution of hardware for picture and video media has been remarkable. The trend of multimedia source digitization has accelerated the combining of many multimedia processing techniques, such as storage, editing, and presentation on computers. Accordingly, techniques are expected for new methods of presenting and creating multimedia content. In order to construct an advanced multimedia-processing environment, a huge multimedia database system is required. One of the most important considerations for such a database system is how to reflect the intention of a user in data retrieval. In a typical system, a user retrieves character data by inputting key words. In other database systems, a user uses numeric values on similar images as the key for retrieval. However, in the retrieval of video data with motion or shapes, the retrieval method of a user is not necessarily in agreement with the particular data to be retrieved even when the database system provides graphic user interface (GUI) tools. This paper describes a retrieval technique based on gestures for video images. By using gestures, a user can input spatiotemporal information intuitively. For example, a gesture can be used to input the motion of an object's direction; the motion of the camera; pan and tilt; the shape of an object; to represent the degree of a scale, and so on. Furthermore, it can objectively match keys and indices because both the motion of gestures and the motion of objects in a video can be obtained automatically. We propose the "Interactive Video Database," which enables the creation and presentation of scenes with gestures by replacing the objects that are extracted. By using this technique, a user can create varied and highly realistic video content via intuitive interaction in real time.

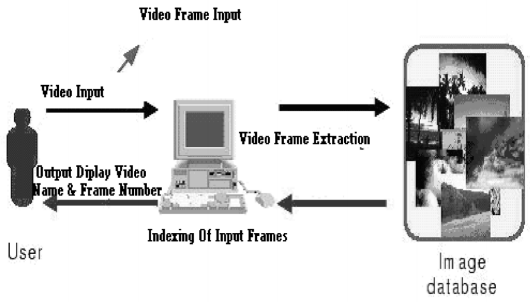


Figure 1 System Architecture

1.2 Video Content Modeling

In order to efficiently describe the video content, we decided to borrow a well-known method used for text document analysis named Latent Semantic Indexing [12]. First we detail the adaptation of LSI to our situation and then propose methods to include multiple features and to improve the robustness of LSI in our particular case, i.e modeling of video shots in a large database. Latent Semantic Indexing is a theory and method for extracting and representing the contextual meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other [13]. In practice, we construct the occurrence matrix A of words into documents. The singular value decomposition of A gives transformation parameters to a singular space where projected documents can efficiently be compared.

For video content analysis, a corpus does not naturally exist, however one can be obtained thanks to vector quantification technics. In [14], we presented an approach on single video sequences that relies on k-means clustering to create a corpus of frame-regions. Basically, key-frames are segmented into regions [15] and each region is represented by a set of features like color histogram and Gabor's energies. They are then mapped into a codebook, obtained with the k-means algorithm, to construct the cooccurrence matrix A of codebook elements in video keyframes.

Thus each frame is represented by the occurrence of codebook terms. LSI is then applied to the matrix A and provides projection parameters U into a singular space where frame vectors are projected to be indexed and compared. This can be extended to model a set of video sequences; the set can be seen as a unique video where keyframes are the representative frames of shots.

Mathematical operations are finally conducted in the following manner:

- First a codebook of frame – tregions is created on a set of training videos,
- The co-occurrence matrix is constructed:

Let A of size M by N be the co-occurrence matrix of ?M centroids (defining a codebook) into N- key frames (representing the video database).its value at cell (i,j) corresponding to the n8mber of times the region I appears in the frame j.

- Next, it is analyzed through LSA:
  The SVD decomposition gives $A = USV^i$ where
  $$\mathbf{UU^i = VV^i = I,\ L = min\ (M,N)}$$
  $$\mathbf{S \approx diag\ (\sigma_1,\ldots\ldots\ \sigma_L\ ),\ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_L}$$
  Then A is approximated by truncating U and V matrices to keep k factors in S corresponding to the highest singular values.
  $$\mathbf{\hat{A} = U_k S_k V_k^t\ with\ S_k = diag(\sigma_1,\ldots\ldots\ \sigma_k)}$$
- Finally, indexing of a context of A noted c(j) and a new context q is realized as follows:
  $$P_{C(j)} = \text{row j of VS}$$
  $$\mathbf{P_q = q^t\ U_k}$$
- And to retrieve the context q in a database containing indexed contexts P_j, the cosine measure m_c is used to compare elements.
  $$m_c(P_j,q) = P_q.P_j\ /\ \|P_q\|.\|P_j\|$$
  the most similar elements to the query are those with the highest value of m_c.

The number of singular values kept for the projection drives the LSA performance. On one hand too many factors are kept, the noise will remain and the detection of synonyms and polysemy of visual terms will fail. On the other hand if too few factors are kept, important information will be lost degrading performances. Unfortunately no solution has yet been found and only experiments allows to find the appropriate factor number.

In the particular situation of video content, many features can be extracted. Three methods were evaluated to consider the multiple features. They are combined at the origin, before the creation of the codebook, or independent codebooks are merged to create a single occurrence matrix, or the LSI is applied to each feature and the similarity measure is modified to combine outputs from each singular space. We retained that equivalent performances were obtained when features were combined just before or after LSI. The latter solution being the most flexible is kept fro our task. Indeed features added without the need to do all computation tasks again.

Contrary to the modeling of a single video content, LSI does not reveal as per formant for many videos. The occurrence information in each frame is too weak compared to approximations inherent to the use of the codebook and this effect is further emphasized when many videos are implied. To compensate for codebook instability, we match a region to its k – nearest elements in the codebook. This one to many relationship allows to inject more occurrence information for each key-frame and to deal with the sub-optimality of the codebook. E observe a real improvement when looking for similar frames in the database .

2. THE PROPOSED APPROACH

2.1 Input Video File Processing

The human visual system appears to be capable of temporally integrating information in a video sequence in such a way that the perceived spatial resolution of a sequence appears much higher than the spatial resolution of an individual frame. While the mechanisms in the human visual system that do this are unknown, the effect is not too surprising given that temporally adjacent frames in a video sequence contain slightly different, but unique, information. By Extracting both the spatial and temporal information present in a short image sequence to create a single high-resolution video frame. A novel observation model based on motion compensated sub sampling is proposed for a video sequence. In this Video File processing applications the histogram mechanism is proposed. Which utilize the motion histogram for video retrieval, clustering and classification? The video database consists of 20 video sequences taken from the AVI, MPEG-7 test set. The videos include different videos (e.g., file downloading, Earth Movement), Folder Copy etc. The duration of each video segment varies from 1 to 5 s and there are a total of 50 frames. We expect various types of motion content in the different videos and therefore, they form a suitable data set in which to test the proposed motion histogram.

2. 2 Splitting Of Frame

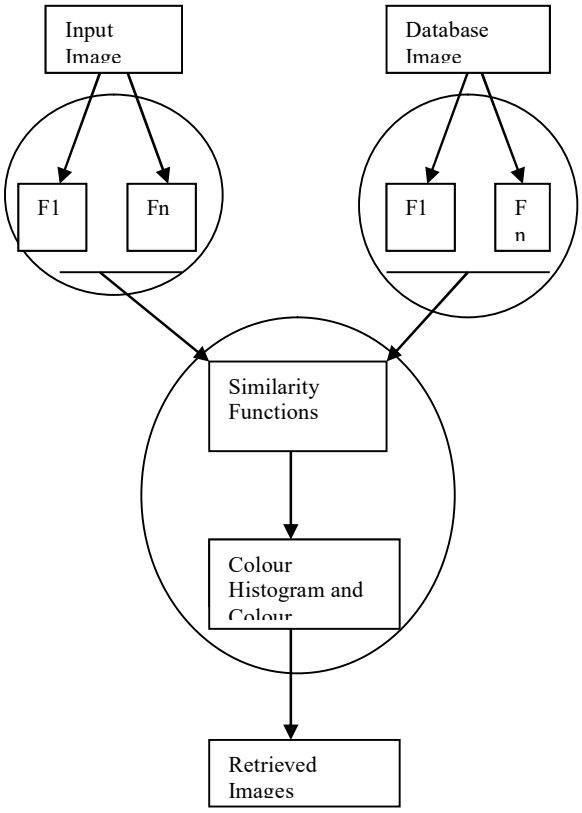
Splitting of frame is the key process and is a simple however effective form of summarizing a long video sequence. The number of key frames used to abstract a shot should be compliant to visual content complexity within the shot and the placement of key frames should represent most salient visual content. Motion is the more salient feature in presenting actions or events in video and, thus, should be the feature to determine key frames. By implementing the model motion patterns in video and a scheme to extract key frames based on this model. The frames at the turning point of the motion acceleration and motion deceleration are selected as key frames. The key-frame selection process is threshold free and fast and the extracted key frames are representative.

2.3 Indexing Of Frame

In order to for the system to fetch the results at a faster rate, there are techniques like catching the most common search or search results and one among them is the Indexing. For any data retrieval the indexing is treated as an important aspect. The concept of indexing is not new to the world, this is used to catalogue the books and other articles in the libraries and book stores where the books are placed in an order and the reader can go to the right place without going through the whole books. As mentioned earlier there are different types of systems that fetch work based on different algorithms and the retrieved image can be the same query image or a part of the query image or even objects in the image. Sometimes the query can be “retrieve ten most similar images to the given image” which can be called as the Nearest-neighbour. In the existing systems the performance is proportional to the database size. The efficiency of the system should depend on the number of similar images rather the total number of images in the database. [4]

There are two ways of indexing an image in the multimedia databases. One is based on the image features and the other coefficients of the images. The proposed system uses the Indexing frames of images and objects there in are manually

identified and described in terms of what they are and what they represent, and in indexing certain features of the images, such as color, are automatically identified and extracted. first method for indexing where the features of the images are calculated and stored in the database along with the image, features like pixel measures, colour histograms etc. These stored values are called Feature Vectors. The same is calculated for the query image. If two images are said to be similar then their feature vectors are similar. The difference between the features vectors finds the distance between the colours. (5)



Where F1 – Fn - Feature Values or vectors from 1 to n
Here feature can be of colour values like the Red, Green and Blue or the values of the generated histogram.

Figure 2.3.1 Indexing of Video Frames

2. 4 Histogram Search

Histogram search algorithms [4], [18] characterize an image by its color distribution or histogram. A histogram is nothing but a graph that represents all the colors and the level of their occurrence in an image irrespective of the type of the image. Few basic properties about an image can be obtained from using a Histogram. It can be used to set a threshold for screening the images. The shape and the concentration of the colors in the histogram will be the same for similar objects even though they are of different colors. Identifying objects in a grey scale image is the easiest one as the histogram is almost similar as the objects have the same colors for same objects. In order for identifying the objects in the images or generating the histogram the system has to obtain the array values.

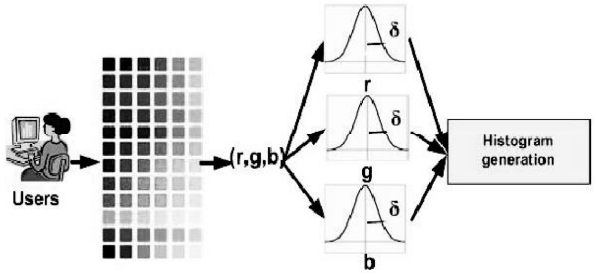


Figure 2.4.1 Color histogram generation process

In general any image contains useful and unwanted information. The system has to differentiate between the both. Consider the below image where the person reading a book is the useful information and the background, people and the market is the unwanted data. The system has to group together the repeated pattern to identify the objects in the image. For example below is given the array for the part of the shirt and this pattern is repeated again

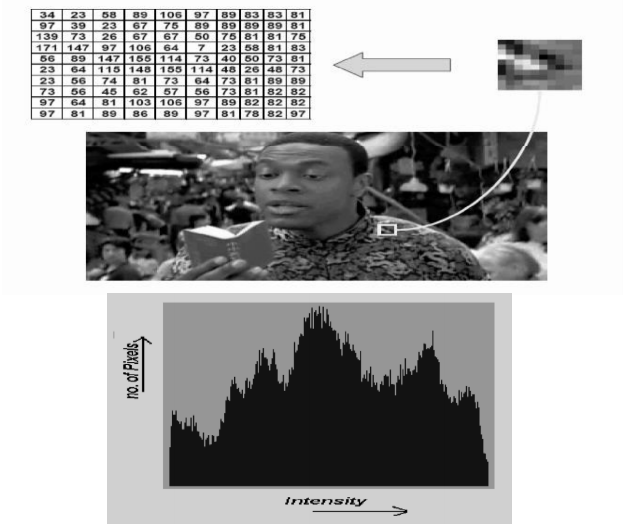


Figure 2.4.2 (Source: Tang, 2007)

Consider the above image where the small part of the person's) shirt is enlarged and the respective representation in the form of the array is given.

The basic concept behind the histogram generation is simple. Each pixel in the image is scanned and the respective color or intensity value is obtained for the pixel.

iColor = (16 * p1[0]) + p1[1] * 4 + p1[2];

Then a graph is generated with total number of pixels against the pixel intensity. An array variable is chosen to store the different intensities and the counter increases for each repeated intensity counting the total number of occurrences of that particular color or intensity.

iHistoArr[iColor] = iHistoArr[iColor] + 1;

An image histogram shows the distribution of pixel intensities within an image. Figure 2.4.3 is an example of an image histogram with amplitude (or color) on the horizontal axis and pixel count on the vertical axis.

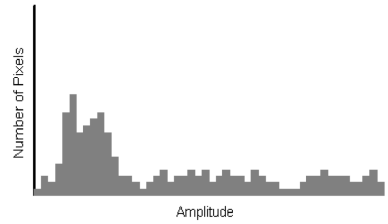


figure :2.4.3 An example of image histogram

2.4.3 Calculation of Image Histogram:
Calculating an image histogram on a sequential device with single thread of execution is fairly easy:

```
for(int i = 0; i < BIN_COUNT; i++)
result[i] = 0;
for(int i = 0; i < dataN; i++)
result[data[i]]++;
```

Listing 1. Histogram calculation on a single-threaded device. (pseudocode).Distribution of the computation process between multiple execution threads is possible. It amounts to :
1) Subdivision of the input data array between execution threads
2) Processing of the sub-arrays by each dedicated execution thread and storing the resulting to a certain number of sub histograms. In some cases it may also be possible to reduce the number of histograms by using atomic operations, but resolving collisions between threads may turn out to be more expensive.
3) Finally all the sub-histograms need to be merged into a single histogram.
When adapting this algorithm to the GPU several constraints should be kept in mind:
Access to the data[ ] array is sequential, but access to result[ ] array is data dependent (random). Due to inherent performance difference between shared and device memory, especially on random patterns, shared memory is the most optimal storage for the result[] array.

On G8x hardware, the total size of the shared memory variables is limited by 16KB.

A single thread block should contain 128-256 threads for efficient execution.
G8x hardware does not have native support for atomic shared memory operations.

3 EXPERIMENTAL RESULTS

Different experiments were performed, one for assessing computational performance, and one assessing robustness with respect to validate the methods we have described, we implemented the components of the video frame based retrieval system and tested with a general purpose image database including about 100 videos. The table given below shows the time taken for Splitting Number of frames from the image database.

Number Of Frames Splitted	Search Time
100	2 Seconds
1000	5 Seconds
10000	14 Seconds

Table 3.1 Average Search Time

These images are stored in JPEG format with size 384 _ 256 or 256 _ 384. The system was written in the VB programming language and compiled on Windows platform. In this section, we describe the training concepts and show indexing results. The table given below shows the number of frames spitted from the videos.

Number Of Frames	Videos
13116	Cartoon
27956	Tennis
71379	News
13254	Movie Song

Table 3.2 Average Search Time For a particular frame



Fig. 3.1. The model input frames of different types of test videos

4. CONCLUSIONS

In this Research, the need for an efficient video based indexing & Video Frame Based retrieval system is identified and the problems with the existing systems are discussed. In an introduction, we have to discuss the video content based model. A new system is proposed for different types of Video Frames and the structure of the Frames and the way they are indexed and stored inside the database are discussed. Different types of images can be processed using the system. The images in the database are compared with the properties of the Video Frame based on color histogram and appropriate results are displayed to the user. In this search histogram we have to discuss the calculation of histogram and also give the best example of image histogram. The architecture is designed by studying different existing systems and current research areas. Research in the fields of image processing, segmentation, edge detection, pattern recognition and more is performed to design the system.

5. REFERENCES

[1] Nevenka Dimitrova, Hong-Jiang Zhang, Behzad Shahraray, Ibrahim Sezan, Thomas Huang, and Avideh Zakhor, “Applications of video-content analysis and retrieval,” IEEE Multimedia Magazine, vol. 9, no. 3, pp. 42 – 55, July 2002.

[2]D.Saravanan, ”Improved image retrieval using image noise removal technique” Pak journal of biotechnology, Vol 14(Special Issue-II-2017), Pages 360-362, Dec- 2017.

[3]D.Saravanan, “Clustering the irregularity events in intelligence surrounding systems” Int. Journal of pure and applied mathematics, Vol. 119, No.12(2018), Pages 15025-15035, May-2018 (Special Issues) , ISSN:1311-8080.

[4] R. Benmokhtar and B. Huet,"Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content," International MultiMedia Modeling Conference, vol. 4351, pp. 196–205, Singapore, 2007.

[5] D.K. Park, Y.S. Jeon and C.S. Won, "Efficient use of local edge histogram descriptor,"ACM Workshops on Multimedia, pp. 51–54, USA, 2000.

[6] M. Rautiainen and T. Seppanen,"Comparison of visual features and fusion techniques in automatic detection of concepts from news video," IEEE International Conference on Multimedia & Expo, The Netherlands, 2005.

[7] E. Allwein, R. Schapire, and Y. Singer, "Reducing multiclass to binary : A unifying approach for margin classifiers." Journal of Machine Learning Research, vol. 1, pp. 113–141, 2000.

[8] Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T.(2004). Probabilistic Author-Topic Models for Information Discovery. The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington.

[9]D.Saravanan," Effective video Content Retrieval using image attributes", EAI Endorsed Transactions on energy Web and Information Technologies, Vol5, Issues18, e8, Pages 1-5. June 2018.

[10]D.Saravanan," Efficient Video indexing and retrieval using hierarchical clustering techniques", Advances in Intelligence systems and computing, Volume 712, Pages 1-8, ISBN:978-981-10-8227,6, Nov-2018.

[11] Tardini, G. Grana C., Marchi, R., Cucchiara, R., (2005). Shot Detection and Motion Analysis for Automatic MPEG-7 Annotation of Sports Videos. In 13th International Conference on Image Analysis and Processing.

[12] Witten, I. and Frank, E. (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005

[13] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990.

[14] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. Discourse Processes, 25:259–284, 1998.

[15] Fabrice Souvannavong, Bernard Merialdo, and Benoˆ Huet. Video content modeling with latent semantic analysis. In Third International Workshop on Content-Based Multimedia Indexing, 2003.

[16] P. Felzenszwalb and D. Huttenlocher. Efficiently computing a good segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 98–104, 1998.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Solving Data Mining Challenge through Link Mining

Ravali Jangam, Thati Sindura
M.Tech Scholor,
Mallareddy College Of Engineering

Mr. Madan Mohan, Asst.prof.
Asst. Professor.
Mallareddy College Of Engineering.

ABSTRACT

A key challenge for data mining is tackling the problem of mining richly structured datasets, where the objects are linked in some way. Links among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical mod-els. Recently there has been a surge of interest in this area, fueled largely by interest in web and hypertext mining, but also by interest in mining social networks, security and law enforcement data, bibliographic citations and epidemiological records.

1. INTRODUCTION

Traditional data mining tasks such as association rule mining, market basket analysis and cluster analysis commonly attempt to find patterns in a dataset characterized by a collection of independent instances of a single relation. This is consistent with the classical statistical inference problem of trying to identify a model given a random sample from a common underlying distribution.

A key challenge for data mining is tackling the problem of mining richly structured, heterogeneous datasets. These datasets are typically multi-relational; they may be described by a rela-tional database, a semi-structured representations such as XML, or using relational or first-order logic. However, the key com-monalities are that the domain consists of a variety of object types and objects can be linked in some manner. In this case, the instances in our dataset are linked in some way, either by an explicit link, such as a URL, or by a constructed link, such as a join operation between tables stored in a database. Naively applying traditional statistical inference procedures, which assume that instances are independent, can lead to in-appropriate conclusions [24]. Care must be taken that poten-tial correlations due to links are handled appropriately. In fact, record linkage is knowledge that should be exploited. Clearly, this is information that can be used to improve the predictive accuracy of the learned models: attributes of linked objects are often correlated and links are more likely to exist between objects that have some commonality.

Link mining is a newly emerging research area that is at the intersection of the work in link analysis [25; 14], hypertext and web mining [3], relational learning and inductive logic pro- gramming [13] and graph mining [8]. Link mining is an instance of multi-relational data mining (in its broadest sense); however, we use the term *link mining* to put an additional emphasis on the links—moving them up to first-class citizens in the data analysis endeavor.

Link mining encompasses a range of tasks including descrip-tive and predictive modeling. Both classification and cluster-ing in linked relational domains require new data mining algo-rithms. But with the introduction of links, new tasks also come to light. Examples include predicting the numbers of links, predicting the type of link between two objects, inferring the existence of a link, inferring the identity of an object, finding co-references, and discovering subgraph patterns. We define these tasks and describe them in more detail in Section 3.

2. BACKGROUND

Probably the most famous example of exploiting link structure is the use of links to improve information retrieval results. Both the well known page rank measure [35] and hubs and authority scores [27] are based on the link structure of the web. These algorithms are based on the citation relation between web pages. Recently, many algorithms have been proposed which examine other relations, for example, Dean and Henzinger [9] proposed an algorithm based on co-citations to find related web pages, or finer-grained representation of the web pages [5]. Richardson and Domingos [40] combined content and link information with a relevance model to improve performance.

A closely related line of work is hypertext and web page classification. This work has its roots in the information retrieval (IR) community. A hypertext collection has a rich structure that should be exploited to improve classification accuracy. In addition to words, hypertext has both incoming and outgoing links. Traditional IR document models do not make full use of the link structure of hypertext. In the web page classifi-cation problem, the web is viewed as a large directed graph. Our objective is to label the category of a web page, based on features of the current page and features of linked neighbors. With the use of linkage information, such as anchor text and neighboring text around each incoming link, better categoriza- tion results can be achieved. Chakrabarti et al. [4] proposed a probabilistic model to utilize both text and linkage information to classify a database of patents and a small web collection. They showed that naively incorporating words from neighbor-ing pages reduces performance, while incorporating category information, such as hierarchical category prefixes, improves performance. Oh et al. [34] reported similar results on a col-

lection of encyclopedia articles: simply incorporating words from neighboring documents was not helpful, while making use of the predicted class of neighboring documents was helpful. These results indicate that simply assuming that link documents are on the same topic, and incorporating the features of linked neighbors, is not generally effective.

Another approach to hypertext and link mining combines techniques from inductive logic programming with statistical learn-ing algorithms to construct features from related documents. A pioneering example is the work of Slattery and Craven [43]. They proposed a model which goes beyond using words in a hypertext document making use of anchor text, neighboring text, capitalized words and alphanumeric words. Using these statistical features and a relational rule learner based on FOIL [39], they proposed a combined model for text classification. Popescul et al. [38] also combined a relational learner with a logistic regression model to improve accuracy for document mining.

Other approaches to link mining identify certain types of hypertext regularities such as encyclopedic regularity (in which linked objects typically have the same class) and co-citation regularity (in which linked objects do not share the same class, but objects that are cited by the same object tend to have the same class). Yang et al. [48] gave an in-depth investigation of the validity of these regularities across several datasets and using a range of classifiers. They found that the usefulness of the regularities varied, depending on both the dataset and the classifier being used.

Another link mining task that has received increasing attention is the identification of communities or groups, based on link structure. Gibson et al. [20] gave a survey of work in discovering Web communities. Kubica et al. [29] proposed a probabilistic model for link detection and modeling groups that makes use of demographic information and linkage information to infer group membership.

Social and collaborative filtering has also been a focus of research that can be viewed as link mining. Kautz et al. [26] constructed social networks from Internet data and used the networks to guide users to experts who can answer their questions. Domingos and Richarson [12] modeled the potential value of a customer, based on their network connections.

Others have proposed generative probabilistic models for linked data. Cohn and Hofmann [7] proposed a probabilistic model for hypertext content and links. We also proposed a generative model for relational data, both content and links [17]. How-ever, depending on the task, predictive models may be more appropriate. Examples of predictive modeling in relational do-mains include [44], [38], and [31].

3. LINK MINING TASKS

As mentioned in the introduction, link mining puts a new twist on some classic data mining tasks, and also poses new problems. Here we provide a (non-exhaustive) list of possible tasks. We illustrate each of them using the following domains as mo-tivations:

Web page collection: In a web page collection, the objects are web pages, and links are in-links, out-links and co-citation links (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor text.

Bibliographic domain: In a bibliographic domain, the objects include papers, authors, institutions, journals and conferences. Links include the paper citations, authorship and co-authorship, affiliations, and the appears-in relation between a paper and a journal or conference.

Epidemiological Studies: In an epidemiology domain, the objects include patients, people they have come in contact with, and disease strains. Links represent contacts between people and which disease strain a person is infected with.

3.1 Link-Based Classification

The most straightforward upgrading of a classic data mining task to linked domains is *link-based classification*. In link-based classification, we are interested in predicting the category of an object, based not just on its attributes, but on the links it participates in, and on attributes of objects linked by some path of edges.

An example of link-based classification that has received a fair amount of attention is web-page classification. In this prob-lem, the goal is predict the category of a web page based on words on the page, links between pages, anchor text and other attributes of the pages and the links. In the bibliographic do-main, an example of link-based classification is predicting the category of a paper, based on its citations, the papers that cite it, and co- citations (papers that are cited with this paper). In the epidemiology domain, an example is the task of predicting the disease type based on characteristics of the people (note the arbitrary possible prediction direction) or predicting the person's age, based on the disease they are infected with and the ages of the people they have been in contact with.

3.2 Link-based Cluster Analysis

The goal in cluster analysis is to find naturally occurring sub-classes. This is done by segmenting the data into groups, where objects in a group are similar to each other and are very dissimilar from objects in different groups. Unlike classification, clustering is unsupervised and can be applied to discover *hidden patterns* from data. This makes it an ideal technique for applications such as scientific data exploration, information retrieval, computational biology, web log analysis, criminal analysis and many others.

There has been extensive research work on clustering in areas such as pattern recognition, statistics and machine learning. Hierarchical agglomerative clustering (HAC) and k-means are two of the most common clustering algorithms. Probabilistic model-based clustering is gaining increasing popularity [21; 45; 29]. All of these algorithms assume that each object is described by a fixed length attribute-value vector.

In the case of clustering linked data, even the definition of an element in a cluster is open to interpretation. We can clus-ter individual objects, collections of linked objects, or some other subgraph of the original. How do we compare the simi-larity of two of these elements or subgraphs, with potentially different structures? As this may necessitate tests for graph-isomorphism, things will quickly become intractable. There has been surprisingly little work done on this type of link min-ing. Subdue [8] is the earliest line of research in this area. More recent approaches have been focused on efficiently find-ing frequently occurring patterns [30; 23]; these are largely inspired by the apriori algorithm [1] for mining frequently oc-

curring patterns. One very interesting new approach is ANF [36], which attempts to compress a graph by approximating the neighborhood function for each node.

Examples of clustering in web page collections range from finding hubs (pages that point to lots of pages of the same category) to identifying mirror sites. Examples of clustering in the bibliographic domain include finding groups of authors that commonly publish together, and discovering research areas, based on common citations and common publication venues and discovering. An example of clustering in the epidemiol-ogy domain is finding patients with similar sets of contacts or diseases with similar transmission patterns.

Next, we turn to some more specific tasks that arise in link mining. These can often be seen as special cases of link- based classification or link-based cluster analysis.

3.3 Identifying Link Type

There is a wide range of tasks related to predicting the existence of links. One of the simplest is predicting the type of link between two entities. For example, we may be trying to predict whether two people who know each other are family members, coworkers, or acquaintances, or whether there is an adviser–advisee relationship between two coauthors.

The link type may be modeled in different ways. In some instances, the link type may simply be an attribute of the link. In this case, we may know the existence of a link between two en- tities, and we are simply interested in predicting its type. In our first example, perhaps we know there is some connection be- tween two people, and we must predict whether it is a familial relation, a coworker relation or acquaintance relation. In other instances, there may be different kinds of links. These may be different potential relationships between entities; in the second example, there are two possible relationships: a co-author re- lationship and an adviser–advisee relationship. We may want to make inferences about the existence of one kind of link, having observed another type of link.

A closely related task is predicting the *purpose* of a link. In a web page collection, the links between pages occur for different reasons. At the coarsest grain, links may be for navigational purposes or for advertising; it may be quite useful to distinguish between the two. The links may also indicate different relationships; the purpose of a link may be to refer to a professor's students, a student's friends, or a course's assignments.

3.4 Predicting Link Strength

Links may also have weights associated with them. In a web page collection, the weight may be interpreted as the authoritativeness of the incoming link, or its page rank. In an epidemiological domain, the strength of a link between people may be an indication of the length of their exposure.

3.5 Link Cardinality

There are many practical inferences that involve predicting the number of links between objects. The number of links is often a proxy for some more meaningful property whose semantics depend on the particular domain:

> In a bibliographic domain, predicting the number of citations of a paper is an indication of the impact of a paper— papers with more citations are more likely to be seminal.
>
> In a web collection, predicting the number of links to a page is an indication of its authoritativeness; predicting the number of links from a page is an indication that the page is a hub. The page rank measure is also clearly related to the number of links.
>
> In an epidemiological setting, predicting the number of links between a patient and people with whom they have been in contact (their contacts) is an indication of the potential for disease transmission; predicting the number of links between a particular disease strain and people infected by it is an indication of the strain's virulence.

Note that link counts can be generalized to paths. A count of the number of paths between two objects may be significant.

3.6 Record Linkage

Another important concept in link mining is identity uncertainty [41; 37; 2]. In many practical problems, such as information extraction, duplication elimination and citation matching, objects may not have unique identifiers. The challenge is to determine when two similar-looking items in fact refer to the same object. This problem has been studied in statistics under the umbrella of record linkage [46; 47]; it has also been studied in the database community for the task of duplicate elimination [42].

In the link mining setting, it is important to take into account not just the similarity of objects based on their attributes, but also based on their links. In the bibliographic setting, this means taking into account the citations of a paper; note that as matches are identified, new matches may become apparent.

4. STATISTICAL MODELS FOR LINK MINING

Given the above collection of tasks, there are some unique challenges to applying statical modeling techniques. Here, we identify several; see also other papers in this volume, and papers in several recent workshops on learning statistical models from relational data [18; 19].

4.1 Logical vs. Statistical Dependences

The first challenge in link mining and multi-relational data mining is coherently handling two different types of dependence structures:

> **link structure** - the logical relationships between objects
>
> **probabilistic dependency** - the statistical relationship between attributes of objects.

Typically we limit the probabilistic dependence to be among objects that are logically related.

In learning statistical models for multi-relational data, we must not only search over probabilistic dependencies, as is stan-dard in any type of statistical model selection problem, but potentially we must search over the different possible logical relationships between objects. This search over logical rela-tionships has been a focus of research in inductive logic pro-gramming, and the methods and machinery developed in this community should be used to tackle this problem.

4.2 Feature Construction

A second challenge is feature construction in the multi-relational setting. The attributes of an object provide a basic description of the object. Traditional classification algorithms are based on these types of object features. In a link-based approach, it may also make sense to use attributes of linked objects. Fur-ther, if the links themselves have attributes, these may also be used. This is the idea behind propositionalization [15; 28]. However, as others have noted, simply flattening the relational neighborhood around an object can be problematic. Several have noted that in hypertext domains, simply including words from neighboring pages degrades classification performance [4; 34]. A further issue is how to deal appropriately with rela-tionships that are not one-to-one. In this case, it may be appro-priate to compute *aggregate* features over the set of related ob-jects. We have found this works well for learning probabilistic relational models [16], but this approach may not always be appropriate.

4.3 Collective Classification

A third challenge is classification using a learned model. A learned link-based model specifies a distribution over link and content attributes, which may be correlated based on the links between them. Intuitively, for linked objects, updating the category of one object can influence our inference about the categories of its linked neighbors. This requires a more com-plex classification algorithm than for a propositional learner. Iterative classification algorithms have been proposed for hy-pertext categorization [4; 34] and for relational learning [33; 45; 44]. The general approach of iterative classification has been studied in numerous fields, including relaxation-labeling in computer vision [22], inference in Markov random fields

[6] and loopy belief propagation in Bayesian networks [32]. Some approaches make assumptions about the influence of the neighbor's categories (such as that linked objects have similar categories); we believe it is important to *learn* how the link distribution affects the category. As an example, this allows us to learn the notion of hubs – e.g., a computer science department homepage is likely to point to a lot of professor homepages.

4.4 Effective Use of Unlabeled Data

Recently there has been increased interest in learning using a mix of labeled and unlabeled data. General approaches include semi-supervised learning, co-training and transductive inference. There are some the unique ways in which unlabeled data can be used to improve classification performance in relational domains:

> Just as in the case of the classical machine learning frame-work, in which there are no links among the data, unla-beled data can help us learn the distribution over object descriptions.
>
> Links among the unlabeled data (or test set) can provide information that can help with classification.
>
> Links between the labeled training data and unlabeled (test) data induce dependencies that should not be ignored.

4.5 Link Prediction

A fifth challenge is link discovery, or predicting the existence of links between objects. A range of the tasks that we have described fall under the category of link prediction. A difficulty here is that the prior probability of a link among any set of individuals is typically quite low. While we have had some success with simple probabilistic models of link existence [17], we believe this is an area where there is much research to be done.

A further challenge is the discovery of common relational patterns or subgraphs; some progress has been made in this area [8; 30; 10]; however, this is an inherently difficult problem.

4.6 Object Identity

A final challenge is identity detection. How do we infer aliases, i.e., determine that two objects refer to the same individual? As mentioned earlier, some work has been done in this area by several research communities, but there is a great deal of room for additional work.

Another aspect of this challenge is whether our statistical mod-els refer explicitly to individuals, or only to classes or categories of objects. In many cases, we'd like to model that a connection to a particular object or individual is highly predictive; on the other hand, if we'd like to have our models generalize and be applicable to new, unseen objects, we also have to be able to model with and reason about generic collections of objects.

5. CONCLUSION

There has been a growing interest in learning from linked data, which are described by a graph in which the nodes in the graph are objects and the edges/hyper-edges in the graph are links—or relations—between objects. Tasks include hypertext classification, segmentation, information extraction, searching and information retrieval, discovery of authorities and link discovery. Domains include the world- wide web, bibliographic certifications, criminology and bio- informatics, to name just a few. Learning tasks range from predictive tasks, such as classification, to descriptive tasks, such as the discovery of frequently occurring sub-patterns. We have given a brief summary of some of the work in this area, and some of the challenges in link mining. Link mining is a promising new area where relational learning meets statistical modeling; we believe many new and interesting machine learning research problems lie at the intersection, and it is a research area "whose time has come" [11].

Acknowledgments

This study was supported in part by the Advanced Research and Development Activity (ARDA) under Award Number NMA401-02-1-2018. The views, opinions, and findings contained in this report are those of the author(s) and should not be construed as an official Department of Defense position, policy, or decision unless so designated by other official documentation.

6. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[2] M. Bilenko and R. J. Mooney. On evaluation and training-set construction for duplicate detection. under review.

[3] S. Chakrabarti. *Mining the Web*. Morgan Kaufman, 2002.

[4] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proc of SIGMOD-98*, 1998.

[5] S. Chakrabarti, M. Joshi, and V. Tawde. Enhanced topic distillation using text, markup tags, and hyperlinks. In *Research and Development in Information Retrieval*, pages 208–216, 2001.

[6] R. Chellappa and A. Jain. *Markov random fields: theory and applications*. Academic Press, Boston, 1993.

[7] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems 13*, 2001.

[8] D. Cook and L. Holder. Graph-based data mining. *IEEE Intelligent Systems*, 15(2):32–41, 2000.

[9] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, 31(11–16):1467–1479, 1999.

[10] L. Dehaspe, H. Toivonen, and R. D. King. Find-ing frequent substructures in chemical compounds. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, editors, *4th International Conference on Knowledge Discovery and Data Mining*, pages 30–36. AAAI Press., 1998.

[11] P. Domingos. Prospects and challenges for multi-relational data mining. *SIGKDD Explorations*, 2003. In this volume.

[12] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, 2001.

[13] S. Dzeroski and N. Lavrac, editors. *Relational Data Min-ing*. Kluwer, Berlin, 2001.

[14] R. Feldman. Link analysis: Current state of the art. In *KDD-02 Tutorial*, 2002.

[15] P. A. Flach and N. Lavrac. The role of feature construction in inductive rule learning. In *Proc. of the ICML2000 workshop on Attribute-Value and Relational Learning: crossing the boundaries*, 2000.

[16] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In S. Dzeroski and N. Lavrac, editors, *Relational Data Mining*, pages 307–335. Kluwer, 2001.

[17] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models with link uncertainty. *Journal of Machine Learning Research*, 2002.

[18] L. Getoor and D. Jensen. *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*. AAAI Press, 2000.

[19] L. Getoor and D. Jensen. *Proc. IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*. AAAI Press, 2003.

[20] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[21] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

[22] R. Hummel and S. Zucker. On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):267–287, 1983.

[23] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Principles of Data Mining and Knowledge Discovery*, pages 13–23, 2000.

[24] D. Jensen. Statistical challenges to inductive inference in linked data. In *Seventh International Workshop on Artificial Intelligence and Statistics*, 1999.

[25] D. Jensen and H. Goldberg. *AAAI Fall Symposium on AI and Link Analysis*. AAAI Press, 1998.

[26] H. Kautz, B. Selman, and M. Shah. Referral Web: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[27] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[28] S. Kramer, N. Lavrac, and P. Flach. Propositionalization approaches to relational data mining. In S. Dzeroski and N. Lavrac, editors, *Relational Data Mining*, pages 262— 291. Kluwer, 2001.

[29] J. Kubica, A. Moore, J. Schneider, and Y. Yang. Stochas-tic link and group detection. In *Proc. of AAAI-02*, 2002.

[30] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *ICDM*, pages 313–320, 2001.

[31] Q. Lu and L. Getoor. Link-based classification. In *Proc. of ICML-03*, 2003.

[32] K. Murphy and Y. Weiss. Loopy belief propagation for approximate inference: an empirical study. In *Proc. of UAI-99*. Morgan Kaufman, 1999.

[33] J. Neville and D. Jensen. Iterative classification in relational data. In *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*. AAAI Press, 2000.

[34] H.-J. Oh, S. H. Myaeng, and M.-H. Lee. A practical hypertext categorization method using links and incrementally available class information. In *Proc. of SIGIR-00*, 2000.

[35] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bring order to the web. Technical report, Stanford University, 1998.

[36] C. Palmer, P. Gibbons, and C. Faloutsos. Anf: A fast and scalable toole for data mining in massive graphs. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, 2002.

[37] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*. MIT Press, 2003.

[38] A. Popescul, L. Ungar, S. Lawrence, and D. Pennock. Towards structural logistic regression: Combing relational and statistical learning. In *KDD Workshop on Multi-Relational Data Mining*, 2002.

[39] J. R. Quinlan and R. M. Cameron-Jones. FOIL: A midterm report. In *Proc. of ECML-93*, 1993.

[40] M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

[41] S. Russell. Identity uncertainty. In *Proc. of IFSA-01*, Vancouver, 2001.

[42] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, 2002.

[43] S. Slattery and M. Craven. Combining statistical and relational methods for learning in hypertext domains. In *Proc. of ILP-98*, 1998.

[44] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proc. of UAI-02*, pages 485–492, Edmonton, Canada, 2002.

[45] B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clustering in relational data. In *Proc. of IJCAI-01*, 2001.

[46] W. E. Winkler. Advanced methods for record linkage. Technical report, Statistical Research Division, U.S. Census Bureau, 1994.

[47] W. E. Winkler. Methods for record linkage and bayesian networks. Technical report, Statistical Research Division, U.S. Census Bureau, 1994.

[48] Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 18(2-3):219–241, 2002.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Secured Data - Mobile Cloud Computing

#1Sunman Naika,
#2Lagshetty Anantha Lakshmi,
M.Tech., Scholar(s),
Dept of CSE,
Malla Reddy College of Engineering
Hyderabad

Mr. K.Arun,
Asst.Professor,
Dept. of CSE,
Malla Reddy College of Engineering,
Hyderabad-

Abstract—Cloud computing is an Internet-based computing pattern through which shared resources are provided to devices on-demand. Its an emerging but promising paradigm to integrating mobile devices into cloud computing, and the integration performs in the cloud based hierarchical multi-user data-shared environment. With integrating into cloud computing, security issues such as data confidentiality and user authority may arise in the mobile cloud computing system, and it is concerned as the main constraints to the developments of mobile cloud computing. In order to provide safe and secure operation, a hierarchical access control method u sing modified hierarchical attribute-based encryption (M-HABE) and a modified three-layer structure is proposed in this paper. In a specific mobile cloud computing model, enormous data which may be from all kinds of mobile devices, such as smart phones, functioned phones and PDAs and so on can be controlled and monitored by the system, and the data can be sensitive to unauthorized third party and constraint to legal users as well. The novel scheme mainly focuses on the data processing, storing and accessing, which is designed to ensure the users with legal authorities to get corresponding classified data and to restrict illegal users and unauthorized legal users get access to the data, which makes it extremely suitable for the mobile cloud computing paradigms.

I. INTRODUCTION

With explosive growth of mobile devices including smart phones, PDAs, and tablet computers and the applications installed in them, the mobile-Internet will maintain the development growth trend as 4G communication network is extensively promoted to our lives. What users of the mobile devices and applications need is that mobile-Internet can provide them with the service which is user-friendly, high- speed, and steady. In addition, the security issues of mobile terminals and the Internet access are attached importance to. And as a combination of cloud computing, mobile devices and wireless networks, mobile cloud computing is an emerging but very promising paradigm which brings rich computational resources to mobile users, network operators, as well as cloud computing providers [1] [2] [3]. The flaws of data storing and data computing in mobile-Internet applications can be overcome by mobile cloud computing while the new paradigm can also accomplish cloud based multi-user data sharing, end geographical service limitation, and process real-time tasks efficiently at the same time.

There is no accurate definition of mobile cloud computing, several concepts were proposed, and two most popular schemes can be described as follows:

1) Mobile cloud computing is a kind of scheme which could run an application such as a weather monitor application on remote cloud servers as displayed in Figure 1, while the mobile devices just act like normal PCs except that the mobile devices connect to cloud servers via 3G or 4G while PCs through Internet. And this concept is considered as the most popular definition of mobile cloud computing [4].
2) Taking advantages of leisure resources such as CPU, memory, and storing disks, another model of mobile

cloud computing exploits the mobile devices themselves as resources providers of cloud [5]. And the scheme supports user mobility, and recognizes the potential of mobile clouds to do collective sensing as well-In this paper, we mainly use the first paradigm mentioned above, but the second one inspires us to assume that what if the mobile devices do not provide computing resources or storing resources but sensing data instead?

In fact, most mobile devices are capable to capture some data from the environment nowadays, for example, almost every smart phone are equipped with sensors of proximity, accelerometer, gyroscope, compass, barometer, camera, GPS, microphone [6], etc. Combining the concept of WSN, mobile devices can be regarded as mobile sensors that are able to provide other mobile devices who are users of the mobile cloud services with some sensing information including environment monitoring data, health monitoring data, and so on.

We take a weather monitor application as an example in this paper.

Assuming that a company develops a weather monitor application which aims to share real-time weather information such as temperature, humidity, pictures, and precise location information and so on to other users of the application. And the application utilizes the user-cloud-user model instead of peer-to-peer model so that the users can get classified and demanded information. Another feature of the application is that the users are divided into different hierarchies, depending on which users can get different sensing data, and users with higher privilege level can, of course, get access to more specific and more frequently updated information.

In order to meet what the application requires, security issues of the whole system should not be ignored, among all security issues the most important two security issues in such model be divided into two parts: authority of

Fig. 1. A mobile cloud computing model

users and the confidentiality of sensing data. Those issues can be solved by providing methods of access control [7]. Attribute Based Encryption (ABE) is a recent cryptographic primitive which has been used for access control [8]–[11]. Access control issue deals with providing access to authorized users and preventing unauthorized users to access data. Attaching a list of authorized users to each data is the simplest solution to achieve access control. However, this solution is difficult in the scenario with a large number of users, such as the application mentioned above within the environment of cloud. Public cryptographic scheme is another solution, in which a public/secret key pair is given to each user and encrypt each message with public key of the authorized user, so that only the specific users are able to decrypt it. In the proposed scenario, users with different privilege levels have different rights to access the part of sensing data coming from the mobile devices. Therefore, one same data has to be encrypted into ciphertext once, which ought to be able to be decrypted multiple times by different authorized users.

Based on such application demands, the concept of attribute based encryption is introduced [11]. Senders encrypt message with certain attributes of the authorized receivers. The ABE based access control method uses several tags to mark the attributes that a specific authorized user needs to possess. The users with certain tag sets can get access to the specific encrypted data and decrypt it. Lots of paper [12]–[15] introduced the scheme about the attribute based encryption access control method in the cloud computing. In the mobile loud computing environment, there are tremendous data which needs to be processed and marked with attributions for the convenient attributing access before storing. At the same time, the hierarchical structure of the application users need an authentication center entity to control their attributes. In this paper, a hierarchical access control method using a modified hierarchical attribute-based encryption (M-HABE) and a modified three-layer structure [16] is proposed. Differing from the existing paradigms such as the HABE algorithm and the original three-layer structure, the novel scheme mainly focuses on the data processing, storing and accessing, which is designed to ensure the application users with legal access

authorities to get corresponding sensing data and to restrict illegal users and unauthorized legal users get access to the data, the proposed promising paradigm makes it extremely suitable for the mobile cloud computing based paradigm. What should be emphasized is that the most important highlight of all in the proposed paper can be described as that the modified three-layer structure is designed for solving the security issues illustrated above.

The rest of the paper can be organized as follows. Section 2 points out the security issues in mobile cloud computing paradigm, which can be divided into 2 portions including se-curity issues of cloud computing and security issues of mobile cloud computing. The proposed modified hierarchical attribute based encryption access control method is described in section
3. Section 4 demonstrates how the proposed access control method based on M-HABE applies in a weather application scenario specifically. Conclusions are given in section

II. MOBILE CLOUD COMPUTING SECURITY ISSUES

More and more users are starting to use mobile cloud computing services such as iCloud and OneDrive services because of the poor storage and computation capability of current mobile devices. However, these kind of mobile cloud services are considered to be vulnerable in security and users may lose their stored files or messages such as pictures, documents, contacts, and calendars, what's worse, those information may be stolen by third parties. In September, 2014, Apple admitted that iCloud was compromised by hackers and many pictures of celebrities leaked out [17].

Such leakage event alarmed us that the security issues of mobile cloud should be taken seriously. For solving such security challenges, data authority and data confidentiality should be paid more attention.

Authority of data users: Different authority-level system to get access to sensing data for application users should be established since the paradigm is applied in the hierarchical multi-user shared environment, which also means that the users with higher authority level should get all the data that the users with lower privilege level could get access to, while the lower privilege users cant get the data beyond his/her authority.

Confidentiality of data: Although the cloud services utilized in the scenario are provided by private cloud which is supposed to be secure, it is still necessary to ensure the sensing data protected from malicious third parties that do not belong to the mobile cloud system. Therefore it is important for the system to bring in a secure and efficient encryption scheme.

In this section, we mainly discuss the general cloud computing security issues and mobile cloud computing issues.

A. Security Issues for Cloud Computing

As long as the data is transmitted to cloud, it is utilizing cloud services like IaaS or DaaS, security challenges of which must be overcome since then. There are plenty of research results about cloud security, in conclusion, a secure cloud should at least satisfy 4 basic urges of consumers [12], say availability, confidentiality, data integrity, control.

1) Availability

Cloud providers should offer services that consumers could get and use at any places and any time. There are mainly two methods to enhance availability in cloud, which are virtualization and redundancy. Currently, cloud technology is mainly based virtual machine [13], since cloud providers can provide separated virtualized memory, virtualized storage, and virtualized CPU cycles, so that users can always get them.

Large cloud provider enterprises build data centers in mul-tiple regions all over the world to protect files they store from failing in one particular region and spreading to other regions. For example, Google set three replications for each object stored in it [14], all these redundancy strategies are enhancing the availability for consumers to get whatever they want at any time and any place.

Besides these concerns on availability, don't trust HTTP protocol too much as it is a stateless protocol for

attackers, which may cause unauthorized access to the management interface of cloud infrastructures [13].

2) Confidentiality

Confidentiality has been a huge barrier for cloud providers to popularize cloud to consumers since it comes out. It is understandable that consumers cannot trust the cloud services, after all, nobody knows what will happen to the files, especially important and confidential ones, once they are placed in cloud vendors' hosts.

There basically exist two common approaches in current cloud infrastructures, say physical isolation and encryption. Physical isolation specifically means virtual physical isolation as cloud services are transmitted via public networks. In this context, virtual physical isolation [15], [18] are using VPN and firewalls to secure database [12]. Encrypting vital and confidential data before placing it in cloud infrastructures is another method to enhance confidentiality of cloud. But do not count on that approach too much because novel methods of breaking cryptographic algorithms are discovered [13].

3) Data integrity

Data integrity ensures consumers that their storing data is not modified by others or collapsing owing to system failure. An easy method is making plenty of copies of consumers' files, which is a good but highly- cost way. Besides the method, a ‖cloud security capture application‖ [19] could be in use to show consumers when and where their data was modified or transmitted.

4) Control

It is a sophisticated work to control a cloud system, a controlling work mainly includes deciding what resource could be utilized in what occasions.

In order to own a secure control system, cloud vendors may need a specialized operating system. Virtualization based cloud services make it difficult to overcome defects in se-curity control because of the insufficient control mechanisms that virtualized networks offer. And poor key management procedures of virtualized based cloud services make it worse [13]. Because virtual machines don't have a fixed hardware infrastructure and cloud-based content is often geographically distributed, it is a very tough task to ensure a secure control in cloud.

We conclude the discussion in Table I.

TABLE I

SECURITY ISSUES FOR CLOUD COMPUTING

Security Challenges	Descriptions
Availability	Cloud providers are supposed to guarantee to consumers that they can get and use their data at any places and any time.
Confidentiality	Consumers' data should be kept secret in cloud systems.
Data Integrity	The data stored in cloud systems need a mechanism to ensure their data not lost or modified by unauthorized users.
Control	A secure control system distributes appropriate resources to be utilized in different occasions.

B. Security Issues for Mobile Cloud Computing

Mobile cloud computing model in this paper means that mobile device users run applications on remote cloud servers instead of mobile devices themselves, the paradigm performs almost the same as normal cloud computing with computers except that mobile cloud model connects mobile devices and cloud servers through 3G or 4G while cloud computing paradigm via Internet, therefore, mobile cloud computing inherits the security threats of traditional cloud computing. Whats more, the security issues that are specific to mobile devices such as battery exhaustion attacks
[20] mobile bonnets and targeted attacks
[21] should be concerned as well [1] .

III. MODIFIED HIERARCHICAL ATTRIBUTE-BASED ENCRYPTION (M-HABE)

We take a weather application on mobile devices as a scenario.

As the mobile cloud computing defines [1] [22], there would be so much sensing data from the mobile devices inbursting into the cloud

infrastructures to process and store the data. The sensing data belonging to a mobile cloud computing model can contain information of different hierarchies such as temperature and humanity numbers, the weather changing trend, information update frequency and so on. It is important that the users with lower privilege cannot get access to some information that the higher privilege user can get to, while the higher authority user can get access to all the data that is obtainable for users in lower hierarchical position since different users of the mobile cloud computing system constitute a hierarchical authority system. At the same, all the information should be encrypted appropriately since the data is not supposed to be available for a third party which.

TABLE II
LIST OF THE ACCESS STRUCTURE IN A MOBILE CLOUD COMPUTING MODEL

Information Type	Access Structure
Temperature Value	All users
Weather Horizontal Comparison	Users of level 2, users in the cooperative companyf meteorology company, : : : g
Update Frequency	Users of level 3, users with titles fprofessor, doctor, major journalist, : : : g
Weather Vertical Comparison	Users of level 4, users with designated work numbers f290381, 209378, 98302, : : : g, users doing jobs of fmeteorology researcher, journalist, : : : g
Location Accuracy	Users of level 5, users in specific locations

belong to the system. So a secure and hierarchical access control method should be proposed to apply in the mobile cloud computing system. An example of the access control list in such circumstance is shown in Table II.

As the list suggests, the access structure should meet the following requirements:

a) One encrypted data can be received by several users.
b) Not only precise level descriptions, but users attributes are there in the access structure. For example, weather vertical comparison information can be attainable for users of level 4, some users with working numbers f290381, 209378, 98302, : : : g, and users doing jobs such as meteorology researcher, journalist and so on, among which the users with specific numbers and users doing specific jobs are described as the attributes while the other one is described as an accurate privilege level.
c) The structure of encryption keys should performs just as the hierarchical structure of the mobile cloud computing users. For instance, the way the authentication center within the mobile cloud computing application company distributes the keys to users should just be as how the users privileges show.

In order to accomplish these requirements, a proposed access control method should contain the following features:

a) One ciphertext can be decrypted by several keys.
b) Both precise level description and user attribute should be supported in the access structure of the method.
c) The keys in the authentication center ought to have the same hierarchical structure just as the structure of users privilege levels.

A modified hierarchical attribute-based encryption (M-HABE) access control method applied in mobile cloud com- puting is proposed in this paper, which changes a proposed scheme called hierarchical attribute-based encryption HABE [8], M-HABE combines the hierarchical identity-based en- cryption [9] and the ciphertext-policy attribute-based encryp- tion (CP-ABE) [10] to meet the conditions described above.

A. Related Work

1) Hierarchical identity-based encryption

The concept of Identity Based Encryption (IBE) was pro- posed by Shamir [11] first in 1984, differing from traditional symmetrical encryption system, IBE took arbitrary character strings that can represent the identities of users, such as ID numbers, e-mail addresses, as public keys to encrypt data. One advantage of IBE is that the sender didnt have to search the public keys

information on certificate authority (CA) online, which solved the problem of poor CA performance. The shortage of IBE system was that all users keys were generated by the private key generation (PKG), which would become the bottleneck in the system.

Horwitz [23] proposed the idea of hierarchical IBE (HIBE) in 2002, a user in the higher hierarchical position of the system could create private keys for lower position users with his/her private keys. Which mean that only the first level users private keys need be created by PKG, while lower-level users private keys could be generated and managed by their ancestors. This improved system relieved PKG of great burden and enhanced the system efficiency by authenticating identities and transporting keys within locality area instead of global area.

The public key of a user is described by a set of IDs composed of the public key of father node and the users own ID in the method of G-HIBE [9], the most important feature of the proposal is that the users public key could reflect precise position of the user in the hierarchical structure.

2) Ciphertext-policy attribute-based encryption

Attribute based encryption (ABE) [24] is regarded as the IBE method with an access structure bringing into the ci- phertext or private key, the access structure determines what ciphertext can be obtained by which users.

Two major branches of ABE system are key-policy ABE (KP-ABE) [24] and ciphertext-policy ABE (CP-ABE) [10], the later one is utilized in many paradigms including this proposed paper. The access structure mentioned above in CP-ABE is placed in ciphertext, which means that the data sender can be so initiative that he/she can determine the receiver. Users are described by a set of attributes in CP- ABE, only when the attribute set satisfies the access structure can the user obtains the ciphertext.

The core of the proposed scheme is called modified hi- erarchical attribute-based encryption (M-HABE), which is different from the HABE scheme.

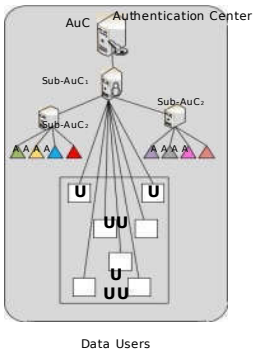
HABE was proposed based on G-HIBE [9] and CP-ABE

[10] by Wang [8] in 2010, it was designed mainly for the usage within an enterprise. We modified the proposal to adapt the scenarios of mobile cloud computing system, which could be illustrated as figure 2, with the aim of making it accommodate to the system based on mobile cloud computing.

As the Figure 2 shows, the proposal consists of an au- thentication center (AuC), Sub-AuCs, and application users. The AuC is responsible for generating and publishing system parameter and the system master key; Sub-AuCs can be

divided into first-level Sub-AuC(Sub-AuC$_i$) and other Sub- AuCs, among which the AuC just need to be in charge of

Fig. 2. M-HABE model



f
o
r
m

o

TABLE III
LIST OF MAJOR KEYS IN HABE

Key Name	Meaning
MK0	Root key, owned by AuC
MK	Master key, owned by Sub-AuC
PK	Public key, owned by Sub-AuC1
PKi	Public key, owned by Sub-AuCs
MKi	Master key, owned by Sub-AuCs
PKu	Public key, owned by users
SKu	Secret key, owned by users
i;u	Secret identity key, owned by users
i;u;a	Secret attribute key, owned by users
PKu	Public key, owned by attributes

users and create their private keys, while other Sub-AuCs take charge of users attributes and create their secret identity keys and secret attribute keys for users.

Each data user shown in the figure possesses a unique ID which is a character string designed to describe the features of internal parties within the system, and so do AuC, Sub-AuCs, and users attributes, especially, the ID of each user contains an integer for describing the privilege level of the user. Additionally, data users also own a set of attributes while other internal parties do not.

B. Key Description

Public key encryption is utilized in the proposed system, the related keys are summarized in Table III.

a) Root key MK0 possessed by AuC is used to create MK for Sub-AuC1.

b) Each Sub-AuC owns a public key PKi and a master key MKi, among which PKi is composed as (PKi 1; IDi) where PKi 1 is the public key of the Sub-AuC's father node, and MKi is also created by the father node. PK is the public key

f (PKi 1; IDi) where PKi is the public key of Sub-AuC that takes charge of the attribute.

C. M-HABE Definition

The M-HABE is composed by the following algorithm-of Sub-AuC1, which can be demonstrated as ID meaning that it is composed by its own IDs. Unlike HABE proposed

by Wang [8], Sub-AuC1 in this paper only needs to take

charge of users, and create their secret keys SKu for them. And other Sub-AuCs have a set of attributes

s:

Setup: Given a security parameter K that is huge enough, AUC will generate a system parameter params and a root master key MK0.

CreateMK: Using system parameter params and their own master keys, AUC or Sub-AuCs can create master keys for lower-level Sub-AuCs.

CreateSK: With its own master key MK and system param-

eter params, Sub-AuC1 creates secret key SKu for each consumer if it is sure that the public key of the user is P

Ku, or there would be no secret key for the user.

CreateUser: Sub-AuCs will create users' secret identity keys SKi;u and secret attribute keys SKi;u;a for them if the Aub-AuC makes sure that the attribute a is in charge of it and the user u satisfies a. And if not there would be no secret identity keys or secret attribute keys.

Encrypt: With R denoting a set of users' IDs, A representing the attribute-based access structure, the pubic keys of all the users that are in R, and the public keys of all the attributes that are in A, the data provider, which is also a data user of the cloud computing in this case, can encrypt the sensing data D into ciphertext C.

RDcrypt: Given the ciphertext C, a data user possessing the precise ID that is in R can decrypt the ciphertext C into

plaintext D with params and the user's secret key SKu.

ADcrypt: Given the ciphertext C, a data user possessing an attribute set fag that satisfies A, which meansthat

the consumer owns at least an attribute key SKi;u;a, can also decrypt the ciphertext C into plaintext D with system parameter params, the user's secret identity

key SKi;u, and the secret attribute key SKi;u;a.

D. M-HABE Construction

ISBN 978-81-939386-1-4

Assuming that IG is a BDH parameter generator, the M-HABE scheme based on bilinear map [9] is constructed by following algorithms:

Note 1: In order to improve the performance of the whole scheme, data user can send Q-tuplei(ti 1) to cloud to explicit kinds of mobile devices, for example, JPEG, WMA, TXT,PNG,W MV,etc. M obiledeviceID is the only sign of the source mobile the greQat computation ability with the aim of helping to comp ute ti $\frac{j}{2}$ e^(Mij; nAQi(j 1)), so that the data user can

decrypt the ciphertext with computing bilinear maps only constant times when running the ADcrypt algorithm.

Note 2: The public key of data consumer u (ID ; IDu) can be combined with the public key of attributes (ID ; : : : ; IDi; IDa) in form of a character string as the input of H1 and HA.

IV. M-HABE ACCESS CONTROL METHOD APPLIED IN CLOUD-BASED SMART GRID

Applying M-HABE, the proposed scheme is illustrated in Figure 3. The whole system works as following steps: All kinds of mobile devices which are installed with the mobile cloud computing based weather application are distributed into different locations all over the country with users. The applications can exploit the sensors installed in the mobile devices to capture the weather data that the applications need, including temperature value, humidity information, atmospheric pressure and so on. The sensing weather data is transported to the layer1 which is a kind of IaaS cloud service provided by the cloud provider [25].

Before sent to layer 2, the sensing weather data is classified by its data model [16] in layer 1 with its excellent ability of computing and storing, the step can be illustrated by figure 4. The data model we present is inspired by the data model proposed in [26], based on which our data model is composed by format, device ID, size, time, value and period. Therefore, a raw data can be expressed as a vector Data hformat; mobiledeviceID; sizeltimelvalue; periodi.

F ormat stands for the basic format of the raw weather data that a specific mobile device produces, there are different kinds of formats depended on different format, and different kinds of mobile devices have different meanings. For example, for a temperature sensor, the value means specific numbers of the temperature, while the humidity sensors can only produce the data with the value attributes that indicate the specific numbers of humidity. A period identification is a time cycle of the sensing data, it is utilized to describe the life period of one specific sensing raw data, and the data will be destroyed once the storing time in cloud of it is beyond the period time.

The sensing weather data is encrypted into ciphertext in layer 2 by M-HABE encryption algorithm using the key in form of (R; A; P Kaja2A; IDu2R),Ł and the cyphertext is sent to layer 3 which is also a kind of IaaS cloud service in cloud. The encryption step can be demonstrated as Figure 5.

The data users of the scheme are in charge of just like Figure 2 indicates. The users can get access to the ciphertexts only if he/she satisfies the requirements of RDcrypt algorithm or ADcrypt algorithm that are described in part III. The decryption procedure is shown in Figure 6.

params=hq; G1; G2; e;^ n; P0; Q0; H1; H2i, among which

device where raw weather data comes from. The size of sensing weather data is defined by the raw weather data itself, which indicates the size of one specific weather data. As for time, as long as a mobile device captures data from the environment where it is in, the time that the sending action occurs will be regarded as the time attribute of the raw sensing data. A value sign represents the most important characteristic of sensing data, the meaning it stands for differs from format to

CONCLUSION

The paper proposed a modified HABE scheme by tak-ing advantages of attributes based encryption (ABE) and hierarchical identity based encryption (HIBE) access control processing. The proposed access control method using M- HABE is designed to be utilized within a hierarchical multi-user data-shared environment, which is extremely suitable for a mobile cloud computing model to protect the data privacy and defend unauthorized access. Compared with the original.

REFERENCES

[1] N. Fernando, S. W. Loke, and W. Rahayu, -Mobile cloud computing: A survey,‖ Future Generation Computer Systems, vol. 29, no. 1, pp. 84–106, 2013.

S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya, -Cloud- based augmentation for mobile devices: motivation, taxonomies, and open challenges,‖ Communications Surveys & Tutorials, IEEE, vol. 16, no. 1, pp. 337–368, 2014. R. Kumar and S. Rajalakshmi, -Mobile cloud computing: Standard approach to protecting and

[2] securing of mobile cloud ecosystems,‖ in Computer Sciences and Applications (CSA), 2013 International Confer-ence on. IEEE, 2013, pp. 663–669.

[3] J. Carolan, S. Gaede, J. Baty, G. Brunette, A. Licht, J. Remmell, L. Tucker, and J. Weise, –Introduction to cloud computing architecture,‖ White Paper, 1st edn. Sun Micro Systems Inc, 2009.

[4] E. E. Marinelli, -Hyrax: cloud computing on mobile devices using mapreduce,‖ DTIC Document, Tech. Rep., 2009.

[5] Q. Han, S. Liang, and H. Zhang, -Mobile cloud sensing, big data, and 5g networks make an intelligent and smart world,‖ Network, IEEE, vol. 29, no. 2, pp. 40–45, 2015.

[6] I. Stojmenovic, -Access control in distributed systems: Merging theory with practice,‖ in Trust, Security and Privacy in Computing and Com-munications (TrustCom), 2011 IEEE 10th International Conference on. IEEE, 2011, pp. 1–2.

G. Wang, Q. Liu, and J. Wu, ―Hierarchical attribute-based encryption for fine-grained

[7]

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Stock Hunt: Predictive Analysis using Machine Learning

Shubhangi Urkude[1]
[1]*Department of Computer Science and Engineering*
Faculty of Science and Technology
The ICFAI Foundation for Higher Education (IFHE)
Hyderabad, Telangana, India – 501203.
ushubhu@ifheindia.org

Niteen Balpande[2]
[2]*Department of Computer Science and Engineering*
Faculty of Science and Technology
The ICFAI Foundation for Higher Education (IFHE)
Hyderabad, Telangana, India – 501203.
cool.viju1721@gmail.com

Abstract--- ***Predicting daily behavior of stocks is a complicated challenge for investors and stockholders. Every day more than 10000 trade companies enlisted in Bombay Stock Exchange (BSE) which is offer an average of 50 million stocks, making an approximate of 50billion Indian rupees in investments as long as stock markets have existed. People who trades stock professionally using variety of analysis methods such as technical, fundamental, quantitative analysis and soon. Analytical methods use a different source ranging from news to price data, but all of the methods aim at predicting the company's stock market prices and their values. They can make decisions on their trading. Nowadays choosing an application based on machine learning algorithms is the right choice. In this paper, linear regression machine learning algorithm is used to predict future stock whether the stock goes up or down.***

Keywords---Bombay stock Exchange (BSE), Stock Holder, Prediction, Linear Regression. Analytically methods, Quantitative Analysis, fundamental Analysis.

I.INTRODUCTION

In financial and corporate world one of the most important thing is stock market trading. With the help of machine learning techniques, the stocks and the market value of any company can be estimated, whether future stock prices are goes up or down. Prediction of stock market is the act of determine the future stock value of the company or other financial instrument traded on an exchange. Prediction of a stock's future price value could yield significant profit.

II.BACKGROUND WORK

In recent decades numerous imperative changes have occurred in the nature of finance markets. The advancement of effective communication and trading aptitude has expanded the extent of choice for investors. Anticipating stock return is a vital financial subject that has pulled in scientist's attention for a long time. It includes presumption that basic information freely available in the past has some prescient connections to the future stock returns (Enke&Thawornwong,2005). To mine such prediction from the accessible data, data mining strategies can be utilized to take out the processed information from this data. Therefore, a few researchers have concentrated on technical analysis and utilizing advanced math and science.

High level of scrutiny has been committed to the field of artificial intelligence and data mining strategies (Wang &chan,2006). Senthamarai Kannan K.et al. [1] proposed different data processing techniques to predict stock exchange. Within the data processing the most issue is historic knowledge that holds the essential memory for predicting the longer-term direction. From the historic stock exchange knowledge, investors discover the hidden pattern of the information that have prognosticative capability within the investment choices. In many statistic predictions, the prediction of stock exchange is considered as difficult task. This paper is able to estimate future stock increases or decrease by the knowledge analysis. The knowledge analysis is additionally a way of prediction. The following strategies are used to analyze and predict the stocks. (i) Typical worth (TP) (ii) Relative Strength Index (RSI) (iii) Bollinger Bands (iv) Moving Average (MA) and (iv) Chaikin Money Flow indicator (CMI). [17]

Huang et al. 2005 [14] proposed Support Vector Machine (SVM) based stock market prediction technique to identify the stock goes up or down. They contrast its performance with Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Elman Back-Propagation Neural Network. They also proposed a combining model by coordinating SVM with other classification method.

Shom Prasad Das et al. [2] proposed Back Propagation Technique and Support Vector Machine Technique have been integrated to predict future prices exchange in the Indian stock market. Zuoquan Zhang et al, 2012 used ε − SVM (ε − support vector machine) to build a stock price prediction model.

Several studies discuss the prophetical ways applied to monetary markets. The sector of Artificial Intelligence, for instance, triggered the event of specialized process tools for the stock market. These tools aim to use market variables to get expected values progressively closer to the particular values of those shares [2]. The Regressive Models, the factitious Neural Networks, the symbolic logic, Genetic Algorithms and Hybrid Models of prediction square measure emphasized.

Koulouriotis, D. E et al. [3] used advanced modeling techniques, such as neural evolutionary, fuzzy and hybrid systems to predict the stock in the market. They proposed Fuzzy Cognitive Maps (FCMs) based stock market model and its effectiveness is tested on real world application data. The FCM training algorithm is showing the accuracy of 60.6% and also it is forecasting the positive and negative stocks.

Nadson S. Timbó. et al. [4] proposed linear regression model for stock exchange prediction and combined with financial indicator to provide support decision making by investors. This model is tested on 30 % test data set and showing success rate of 60.87%.

Binoy B. Nair, N.Mohana Dharini and V.P. Mohandas [5] proposed hybrid tree-adaptive neuro-fuzzy system for predicting stocks. It uses different techniques like technical analysis and call tree. Initially technical analysis is used by stock traders for feature extraction and second call tree for feature choice. By mistreatment technical analysis and call tree that is employed for the reduced dataset is then applied to the adaptational neuro- fuzzy system for next day stock prediction. They tested their projected system on four major international security market. Their experimental results show the projected hybrid system produces abundant higher accuracy when put next to stand- alone call tree based mostly system and ANFIS based system while not feature choice and spatial property reduction. They proposed on top of neuro-fuzzy system.

Stock Prediction Model of A.Subashini, M. Karthikeyan [18] constructed on the statistical computational approach that utilizes Auto-deteriorating Combined Moving Middling (ARIMA). In this paper examined month-wise data that has been found from the S&P Bombay Stock Exchange Complex Index. The dataset extents across a period of six uninterrupted years from 2007 to 2012. Based on the analysis when predicting the future prices of Stock Market safeties, there are numerous theories existing.

A different perception on stock market prediction comes from Random Walk Theory (Malkiel 1973). Stock Market prediction is believed to be terrible where prices are resolute arbitrarily and outdoing the market is infeasible. Random Walk Theory has similar hypothetical reinforcements to Semi-Strong EMH where all public evidence is assumed to be available to everyone. However, Random Walk Theory declares that even with such data, future stock prediction is unproductive.

Debashish Das et al. [15] proposed a solution on reliable stock market prediction by applying hybridization method as data mining techniques and neural network together. It is very effective in unpredictable data like stock market.

Qasem A. Al-Radaideh et al. [16] found that forecasting the stock returns based on historic prices is challenging task, for that he proposed a new model named as CRISP-DM. In which decision is taken based on decision tree classifier.

III. MOTIVATION

Problem of prediction stock market price is that it has the potential to be worth billions and is actively researched by the largest financial corporations and the companies in the world. The variation of stock market is vicious and there are so many complicated financial indicators. However, the advancement in machine learning technology, can help us to find out the most informative indicators to make better prediction. There are so many websites are already present for give the stock market detail of various companies.

IV.PROBLEM STATEMENT

Prediction of the Market task is divides academics and researchers into two groups those who believe that the mechanisms which predict the market and those who believe that whenever new information comes up the stock market can absorb it by correcting and researching itself, because of this there is no space for prediction of stock market. The relation between the activities of market and tendencies of investors found new time scale which operates on updated mood of about 50 million American Facebook users between the periods of 2010-2011. In this paper, predicting of trading volume or stock market price is considered which is similar to the introduced methods.

V.MACHINE LEARNING ALGORITHMS

The stock market price prediction function suggests a few regression models that can be chosen to perform the prediction. This includes

- SVR- Support Vector Regression.
- Linear Regression.

In SVC (Support Vector Classification) can be used to resolve by regression problems. When SVM (Support Vector Machine) is used to resolve regression problems the method is referred as SVR (Support Vector Regression). The model produced by Support Vector Classification depends on the training the data, because the all factor which is useful in SVC can cost of model building which is does not care about training points. Similarly, the model produced by Support Vector Regression SVR only be contingent on the training (Subset) data, as the cost factor for building the model does not contemplate any training data close to the model forecast.

A. LINEAR REGRESSION:

Note that these are the machine learning regression models that were calculated, not all had auspicious results like 100 % accuracy. The outcomes section goes into point of the problems faced with each machine learning regression model and the efforts at the solutions. The following steps to be followed in the Regression Problem:

- Input is mapped by labels.
- Input mapped to large and many data set.
- Use regression when the output can be mapped to many outlets.

The linear regression method is likely to be working fine, but there were some other problems using the machine learning polynomial regression methods, meanwhile the predictions of stock market that are being returned is not the non-linear fitting. In figure 1. by taking non-linear data which is ambiguous and can't be predicted or analyze. But when you apply linear regression on the data it will give you the regression line which is straight line and easy to predict the behavior of stock prices. Where the price details were dependable with the price trends over that same period for relatively short buying periods over a pair of weeks which as seen in Figure 1.

The bond between market trading volume as the dependent factors and the middling price per share (Average)

as the self-determining variable of the regression calculation-[8] is shown in equation 1.

$$Y = \alpha x + \beta \qquad \text{(Equation. 1)}$$

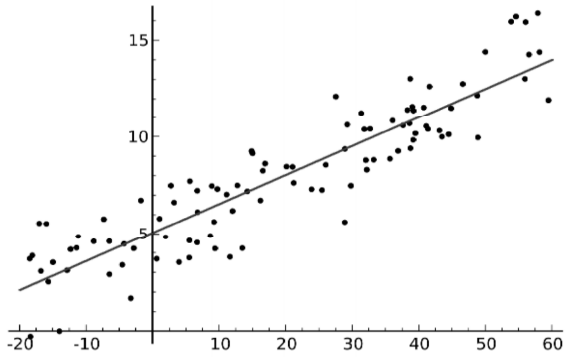


Figure 1: Sample Regression Graph

VI. DATA PROCESSING

A. DATA COLLECTION

The company's stocks which are used in this study are a subdivision of the stocks that are openly traded on the US or any market known as the S&P (Standard & Poor's) 500, which is an index of the 1000 largest companies traded on the NYSE (New York Stock Exchange) or the NASDAQ (Not only US) [5]. In Figure 2, The CRSP (Centre for Research in Security Prices) market dataset that has been provided containing the order of 5000 companies. This data is filtered and loaded into the master data frame to avoid excessive memory usage.

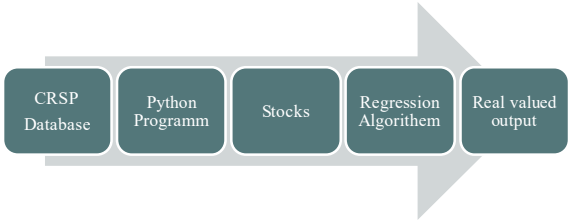


Figure 2: Data Flow for stock market prediction

B. DATA REPRESENTATION

The dataset used in the analysis is collected from the Barchart Stock database. This data set consist of collection of records. Each record contains stock data of a specific day such as shares out from company, closing price of stock, and other features for that day in time. For importing the data from Barchart, API (Application Programming Interface) model need the stock symbol, start date and end date which is given in table 1. Each specific stock is a view of the master data frame that is filtered based on that stock's ticker. This allowed efficient access to stocks of interest and convenient access to date ranges.

Attribute Name	**Domain Value**
Company Symbol	String
Start Date	Integer
End Date	Integer

Table 1: Attribute table

C. PREDICTION THROUGH REGRESSION

The regression process is done through the scikit-learn machine learning library. This is the core for the prediction of stock functionality. There are some more additional steps that must be done so that the stock data can be used into the machine learning regression algorithms and return probable results. In specific, every machine learning training dataset must be normalized before the input matrix is fit to the chosen machine learning regression model.

D. DATA NORMALIZATION

The following process to be followed in the order to be fit into machine learning regression models. First, dates are normally denoted as strings in the format like "YYYY-MM-DD", when it comes to stock database storage. This format must be converted into a single integer in order to be used as a column in the feature stock matrix. This is done by using the date's ordinal value. In python, the columns in the data frame are stored as NumPy datetime objects, which must be converted to Python date and time objects which are in turn converted into integer using the coordinal () which is built-in function for date and time objects in Python. The block diagram for stock prediction is shown in the figure 3.

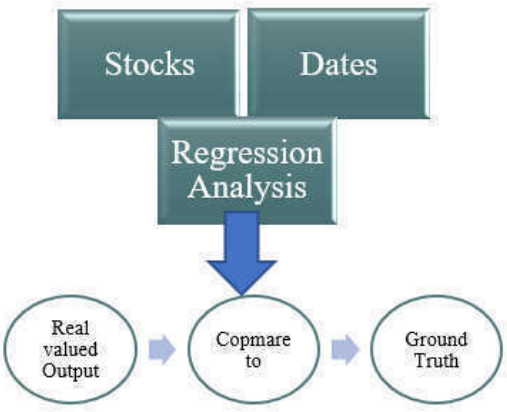


Figure 3: Block Diagram for stock prediction

E. REGRESSION MODEL

The regression model will predict the weather the price will go up or down the next day based on the given the stock name and date. The data collection process is given below:

- Retrieve data.
- Format data.
- Build data sets for training/testing.
- Build graph.
- Train, test, and evaluate model.
- Freeze graph with trained values.

In Figure 3 regression analysis will show the real valued output and error algorithm. Each and every column in the feature stock matrix is well scaled using the scikit learns scale () function (Pyhton library function) from the preprocessing module. The machine learning polynomial regression methods which would return uncertain results since it is documented that the scikit learns (Python library) used in machine learning non-linear regression models, we can say that normally distributed data as the input for feature stock matrices.

VII. RESULT

F. ACCURACY

After finding quantities, slope, error and intercept and put on linear regression market sample data, for testing that how much close and how much accurate the formula can predict the trading stocks market volume to real volume. The predicted stock trading volume is very similar to real values. By calculating the difference among real and predicted values of proposed approach, 60% accuracy is achieved by the model. The formula for accuracy as the given below.

$$\text{Accuracy} = \frac{\text{The number of days that model correctly classified the testing data}}{\text{Total numbers of testing days}}$$

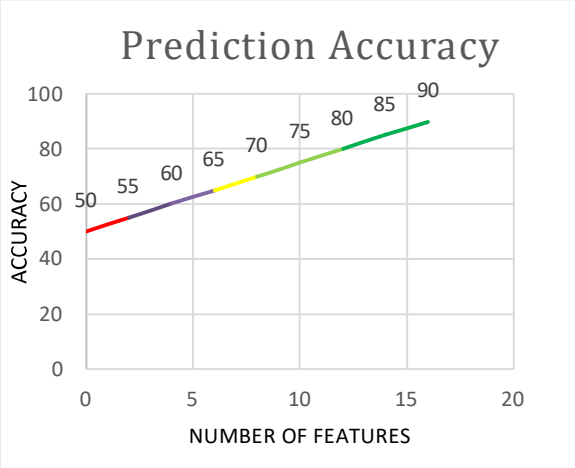


Figure 4: Prediction Accuracy

Summary output of applying the regression analysis has exposed in Table 2. After study all possible algorithms and machine learning techniques there is linear regression analysis is good for predicting the prices whether goes up or down. In table 2 shown the accuracy check for the model. By taking apple and google stock market data.

Stock Symbol	Start date	End date	Predicted	Actual
AAPL	2017-01-01	2017-02-01	Up	Up
AAPL	2017-02-01	2017-04-01	Up	Down
GOGL	2018-02-02	2018-03-15	Down	Down
GOGL	2017-01-01	2017-02-01	Up	Up
AAPL	2014-01-01	2015-01-01	Up	Up
GOGL	2018-03-02	2018-05-15	Down	Up
AAPL	2015-02-01	2016-02-01	Up	Up
GOGL	2018-01-01	2018-10-10	Up	Down
AAPL	2018-02-13	2018-03-18	Down	Down
GOGL	2017-02-01	2017-01-01	Up	Down

Table 2: Accuracy of Regression Model

VIII. DISCUSSION

Markets such as stock market are producing continuously great volume of information needed to breakdown and to produce any predicting pattern in any time. Hence, they are attractive case of using different scientific methods to development and improvement in generating techniques. Each of the used techniques for predicting economic matters has some profits and boundaries of its own which causes to some declining or underwired status. With taking this matter into consideration that our study is a case study on S & P 1000 index to compare with other methods, we checked out 8 most important features for predicting methods. First feature is encoding, which our method is equal to rule training and has high degree in it and it is better than ANOVA and genetic algorithm. Second feature is convenience or accessibility of off-the-shelf software that for this feature it is equal to rule induction, statistical inference and ANOVA and has high degree in it. Third feature is elasticity or capacity of covering diverse types of large scales is equal to algebraic inference and genetic algorithm and has medium degree in it. Fourth feature is autonomy or independence of prior conventions from relations between variables and domain theories is equal to rule induction and algebraic inference and is weak in this feature.

IX.CONCLUSIONS

Each machine learning algorithms are capable of concentrating on particular parts of client's data in electronic shops. This effort brings better and more detailed results to the same parts. In the meantime, in analyzing other parts, due to the lack of analyses, it carries challenges to them.

So, each algorithm is proficient of doing detailed study of some parts of client's market data. To provide inclusive outcomes and analysis, it must be used several integrated and algorithms. This paper of stock market prediction using machine learning, examine different types of methods and clustering algorithms. This paper covers the analyses softness of other algorithms for some clients. The joined data of all algorithms analyses fetches detailed results from customers' behavioral method and its relation with shopping basket as well. So, by using integrated collective data, it can be determined marketing policies and customer satisfaction appropriate to all clients and their orientation which finally lead to increased efficiency and profits.

X.References

[1]. K. Senthamarai Kannan, P. SailapathiSekar, M. Mohamed Sathik and P. Arumugam,"Financial Stock Market Forecast using Data Mining Techniques", International Multi Conference of Engineer and Computer Scientists, Vol 1, pp. 17-19,2010.

[2]. Shom Prasad Das and SudarsanPadhy, "Support Vector Machines for Prediction of Futures Prices in Indian Stock Market", International Journal of Computer Applications, Vol.41– No.3, 2012.

[3]. Koulouriotis, D. E. &Diakoulakis, I. E. & Emiris, D. M "A Fuzzy Cognitive Map based Stock Market Model: Synthesis, Analysis and Experimental Results", IEEE International Fuzzy Systems Conference, pp.465-468, 2001.

[4]. Nadson S. Timbó, Sofiane Labidi, Thiago P. do Nascimento, Milson L. Lima, Gilberto Nunes Neto and Rodrigo C. Matos. ,"Approach Based On Linear Regression For Stock Exchange Prediction – Case Study Of Petr4 Petrobrás" Brazil. International Journal of Artificial Intelligence &Applications, Vol. 7, No. 1, 2016.

[5]. Binoy B Nair, N Mohana Dharini, V P Mohandas, "A Stock Market Trend Prediction System Using a Hybrid Decision Tree-Neuro-Fuzzy System", International conference on Advances in Recent Technologies in Communication and Computing, 2010.

[6]. Enke, D. and Thawornwong, S. "The use of data mining and neural networks for forecasting stock market returns", Expert Systems with Applications, pp.927-940, 2005.

[7]. James m. Poterba and andrew a. Samwick "Stock market fluctuations and consumption behavior", pp. 295-373, 1998.

[8]. Approach and Developing Data Mining Method for Spatial Applications. International Conference on Intelligent Systems & Data Processing, Gharehchopogh, F. S. (2011).

[9]. FS Gharehchopogh and Khaze, S.R, " Data Mining Application for Cyber Space Users Tendency in Blog Writing: A Case Study" International Journal of Computer Applications, Vol. 47, issue no.18, 2012.

[10]. Marc-André Mittermaye, "Forecasting Intraday Stock Price Trends with Text Mining Techniques",Hawaii International Conference on System Sciences, pp. 1-10, 2004.

[11]. Prakash Ramani, Dr. P. D. Murarka,"Stock Market Prediction Using Artificial Neural Network", International Journal of Advanced Research in Computer Science and Software Engineering ,Vol. 3, Issue 4, April 2013.

[12]. John Moody and Mattew Saffell, "Learning to trade via direct reinforcement", IEEE Transactions on Neural Networks, Vol. 12, NO. 4, pp. 875-889,2001.

[13]. Zhai, Y., Hsu, A., &Halgamuge and S. K"Article on Combining news and technical indicators in daily stock price trends prediction", 2007.

[14]. Wei Huang, YoshiteruNakamori andShou-Yang Wang, "Forecasting stock market movement direction with support vector machine",Computers & Operations Research, pp. 2513-2522, 2005.

[15]. Debashish Das and Mohammad Shorif Uddin,"Data Mining And Neural Network Techniques In Stock Market Prediction: A Methodological Review", International Journal of Artificial Intelligence & Applications, Vol.4, No.1, 2013.

[16]. Qasem A. Al-Radaideh, Adel Abu Assaf, and EmanAlnagi, "Predicting Stock Prices Using Data Mining Techniques",The International Arab Conference on Information Technology, pp. 1-8, 2013.

[17]. Dr. Swapna Borde, Austrin F. Dabre, Harsh M. Kamdar and Rahul Y. Pur, "Financial Stock Price Forecast Using Classification", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, 2016.

[18]. A.Subashini and M. Karthikeyan, "Forecasting on Stock Market Time Series Data Using Data Mining Techniques", International Journal of Engineering Science Invention, pp. 06-13, 2018.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A TSM Mechanism to Enhance Node Trustworthiness and Using Asymmetric Key with Ant Intelligence to Secure DSR in Wireless Ad hoc Network.

Veena rani
#*Assistant professor, Department of CSE, MRCE, Maisammaguda*,v4veena.rani@gmail.com
Jakka.Sushma
Assistant professor, Department of CSE, MRCE, Maisammaguda,j.sushu@gmail.com

Abstract: In this paper, a novel approach for node selection for communication in mobile Ad hoc networks is used. Ad hoc networks has limited resources and hence the selection of the trustworthy nodes is one of the most challenging task.. So the TSM mechanism was used and helps to understand the trustworthiness of the node involved in transmitting packets. There are various protocols used in these types of networks and one of the most common protocol being DSR. In this paper, we also introduces a mechanism for securing DSR using asymmetric keys(RSA) and Ant Intelligence(mobile agent).Ant intelligence is a promising technology used in diverse fields of network applications. We found that DSR using ant intelligence/mobile agents with RSA in TSM mechanism can ensure confidentiality to a greater extent.

Keywords **-Ad Hoc Networks, DSR,TSM mechanism, RSA, Ant Agents.**

I. INTRODUCTION:

An ad hoc network can be defined in different ways. It typically refers to any set of networks where all devices have equal status on a network and are free to associate with any other Ad hoc network devices which is in link range. The meaning of Ad hoc is "for this" which means "for this special purpose". The above said networks will need to be based on an open environment and hence it is necessary that we have to select the appropriate nodes for communication. Each node will behave differently in an open environment. In Ad hoc network, each node has its own resources and due to the scarcity of the resources, the behavior of the nodes keeps changing as well. The selfish manner causes the node to be depleting in nature that does not worth in this open environment. So we must be careful of transferring the content from only the well behaving node. Interacting with misbehaving neighbors can be costly for the mobile devices which have limited resources .This problem can be solved by Trust-based solutions.

In this paper, we come up with a mechanism to finding the trust value of node. (The value is calculated based on certain criteria's). However, these solutions usually select the neighbors with the highest trust value. The most trustworthy nodes have to perform

ISBN 978-81-939386-1-4

most of the interactions, which can overload and/or exhaust the resources of these nodes. Furthermore, the trust models tend to evaluate trust values only after an interaction has been accomplished. Thus the trust value of other neighbors may not be evaluated until the most trustworthy node fails. There are different protocols to be used in wireless ad hoc networks. In this paper, we will discuss about the protocol DSR and we will discuss about securing DSR using Asymmetric keys and Ant Intelligence/mobile agents. We depend upon a method known as 'TOPSIS' to choose the shortest path from the different paths available. we move to a secure ad hoc routing approach using the above mentioned technology .

II. LITERATURE SURVEY

David B. Johnson and David A. Maltz proposed a mechanism for the content distribution mobile ad hoc networks. The main aim of the proposed mechanism was to achieve trustworthy node selection along with preservation of mobile node resources.

Djatmiko, M., Boreli, R., Seneviratne, A, & Ries, S discussed about the content distribution and the importance of calculating trustworthiness of nodes before a communication occurs. Here, the author did not mentioned about routing protocol security.

Atul Mishra and A.K.Sharma explored the mobile agent technology utilization in distributed network environment and how it benefits the communication in those environments. This paper has a disadvantage of lack of security as well.

S. Menaka and M.K. Jayanthi introduced intelligent routing in wireless ad hoc networks using the Ant Intelligence technology.

XU Xiao-long and XIONG Jing-Yi,CHENG Chun-Ling Introduced a mechanism for information retrieval by mobile multi agent. In most of the papers, security which is one of the most vital part is not discussed.

In the above papers discussed none of them had explained about security of routing protocols specifically. So, in this proposed system, a methodology is used to improve security to a greater extent. Here, we use TSM method along with asymmetric keys with the usage of ant intelligence to secure DSR routing protocol as well as for ensuring a better secure communication.

III . TSM-Trust Supported Methodology

In a MANET, as we know every node has the capability to transmit packets. Here, for the calculating the trust for each node, observation is done on all nodes which is transmitting the packets. For the above said calculation, we consider the total number of packets they transmit, the packets they receive and also the number of packets they drop due to several reasons. We take into account certain parameters like RREQ(Route Request) , RREP(Route Reply) and DATA(Payload Data). We also consider two criteria's namely success and failure which indicates whether the packet transmission was successful or failure.

ISBN 978-81-939386-1-4

TABLE 1.
PARAMETERS FOR TRUST CALCULATION

Criteria	RREQ	RREP	DATA
Success	$Q_{req}s$	$Q_{rep}s$	$Q_{dat}s$
Failure	$Q_{req}f$	$Q_{rep}f$	$Q_{dat}f$

RREQ stands for route request and RREP stand for route reply .These two parameters are exchanged between the nodes during communication in networks. the payload is the part of transmitted data that is the actual intended message. Headers and metadata are sent only to enable payload delivery. The parameter 'data' refers to the original data. We consider the success rate of 'query request' ie, q_{rs} and it is calculated based on the number of neighboring nodes those who have received the r_{req} successfully from the source node. It is the source node that sends the r_{req}. Another parameter is Q_{rf} which is the failure rate for the query request and is calculated based on number of neighboring nodes those which have not received any query request.

As the initial step, we calculate the request rate, reply rate and data transmission rate of the nodes. We kept a normalized or threshold value for these values and that is -1 to +1. If the values of any above parameters falls below or above the threshold value, that particular node is not consider as genuine or trustworthy. This also means that the failure rate is high and hence we cannot include that particular node for routing . Also, along with the above explained criteria, we also calculate the trust level values of each node during the routing process.

The calculated trust level value is compared with the threshold trust value of the average of all neighboring nodes taking part in routing. If the trust level value calculated falls below the average threshold value , then we can consider that there is a chance that the particular node may drop the packets for that particular transmission and is not suitable for that communication. But there is a chance that this node is used for another communication .Now the next hop is selected completely based on the trust value. Here, first we take all the neighboring nodes and calculate the average trust value. Also, each nodes trust value is also calculated separately. Later this value is compared with the average trust value and the node which has the highest value is chosen. All the remaining nodes are rejected for communication. The threshold value is calculated dynamically at each run. The nodes which is not used as trustworthy node for a particular communication cannot be considered as unfit or untrustworthy node permanently. This is because it may be used as fit node in another transmission.

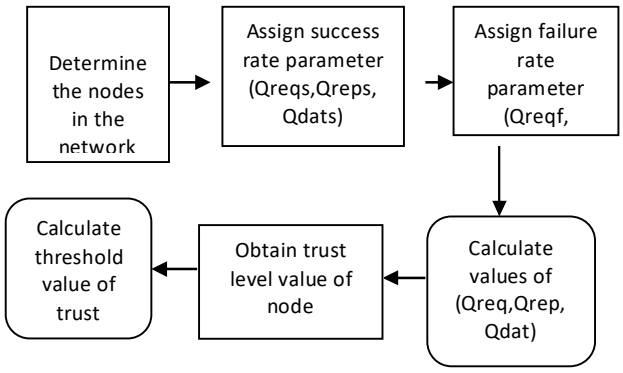


Fig 1: Trust calculation

ISBN 978-81-939386-1-4

Procedure for calculating trust:

Step1: Scan all the nodes from I to n do.

Step2: Initialize the success rate parameters namely, $q_{req}s$, $q_{rep}s$, $q_{dat}s$.

$q_{req}s$ is the route request success rate.

$q_{rep}s$ is the route reply success rate.

$q_{dat}s$ is the data success rate.

step3: Initialize the failure rate parameters namely , $q_{req}f$, $q_{rep}f$, $q_{dat}f$.

$q_{req}f$ is the route request failure rate.

$q_{rep}f$ is the route reply failure rate.

$q_{dat}f$ is the data failure rate.

step4: Initialize the values q_{req}, q_{rep}, q_{dat}

q_{req} denotes the route request.

q_{rep} denotes the route reply.

q_{dat} denotes the payload data.

step5: calculate each values given above.

$q_{req}=\sum_i^n$in $q_{req}s-q_{req}f/q_{req}s+q_{req}f$

$q_{rep}=\sum_i^n q_{rep}s-q_{rep}f/q_{rep}s+q_{rep}f$

$q_{dat}=\sum_i^n q_{dat}s-q_{dat}f/q_{dat}s+q_{dat}f$

step6: calculate trust level value of nodes

$$TL=T(RREQ)*QREQ+T(RREP)*QREP+T(DATA)*QDAT$$

TL is the Trust Level.

step7: For all neighbor nodes j to n do

step8: calculate the threshold value of trust, $TT=\sum_i^n TL/n$

step9:End.

IV.SECURING DATA

Calculating trust value of nodes is one of the way to secure communication. Still, there are lot of problems. Here, we encrypt the data before sending to destination. We use the asymmetric key concept to do this so that security can be achieved to a greater extent. We use RSA algorithm for this purpose. Given below are the steps of encryption and decryption of RSA.

Encryption(RSA)

Sender A does the following

1. Obtain the recipient B's public key(n,e)
2. Represent the plain text message as a positive integer m with 1<m<n
3. Compute the cipher text $c= m^e \bmod n$
4. Send the cipher text c to B

Decrytion(RSA)

Recipient B does the following

1. Uses his private key (n,d) to compute $m=c^d \bmod n$
2. Extracts the plain text from the message representative m.

V. HOW DSR WORKS?

Traditional DSR is basically composed of two parts ie, route discovery and route maintenance. It is purely based on source routing (The node sending a data packet lists in its header the nodes that the packet shall go through) A brief discussion for route discovery and route maintenance is given in the following sub section.

A. Route Discovery: This is the first stage. Route discovery allows any host in the Ad hoc network to dynamically discover a route to any other host in the Ad hoc network(whether directly reachable within wireless transmission range or reachable through one or more intermediate network hops through other hosts). There are two

environment. Mobile agents decide when and where to move. A mobile agent accomplishes a move through data duplication when a mobile agent decides to move, it saves its own state, transport this saved state to the new host, and resumes execution from the saved state.

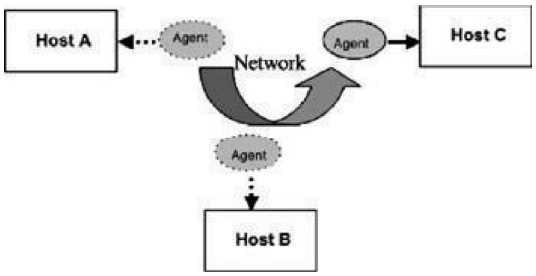


Fig 2: Agent and host Interaction

- ***Ants as Intelligent Agents***

Ants are referred to as dynamic agents . This is because they possess the ability to change their residing locations. These ant agents move out of a particular system to perform a task and may or may not return to the originating node(based on situation). Ants spread intelligence across networks. Also, as they have the mobility feature, The mobility allows them to be created, deployed and terminated without disrupting the network configuration. In fact, external agents that originate at one node keep on changing their locations and may die on any other node shall be referred to as ants. Ants follow the principle of "STIGMERGY". Stigmergy is a form of indirect communication through an environment. The insect ants when move in search of food stimulate a hormone named as pheromones, which attracts other surrounding ants.

VII. ANT INTELLIGENCE

A. DSR with ANT agent: This section of the paper handles how an Ad hoc network may be secured by modifying DSR to use mobile agents.We present TWO types of mobile agents to be used in this routing protocol.

1.Discovery ANT agent.

2.Maintenance ANT agent.

(i) Discovery ANT Agent: Whenever a particular node want to send a data packet to another node(target/destination node) it creates a discovery ANT agent (DA_{ANT}).The DA_{ANT} has a new discovery ID, source node, target node, public key, empty list of passed node. Also, it sets its state to discovery. Next , the sending node broadcast the DA_{ANT} to the nodes in its wireless direct radio transmission range. Whenever a node receives the DA_{ANT}, it checks if it is the target node or not. In case if it is the target node it append tits address to the list of passed nodes, appends the reply which it decrypted with its private key and then changes its states to reply and sends it back to the sending node. If it is not the specified target, it appends its node address to the list of passed nodes in the ANT agent.

(ii) Maintenance ANT agent(MA_{ANT}): The maintenance ANT agent is used whenever there is a truncated hop in a route discovered before. MA_{ANT} is used to notify the sending node with truncated hop to re-discover a route to the target.

It is required to detect and monitor links to neighbors and this is normally done using the hello messages. Periodic broadcasting of hello messages are done by each and every node such that all its neighbors those that are within its transmission range will receive

ISBN 978-81-939386-1-4

this. Normally, the hello messages are transmitted at an interval called hello interval seconds. As the hello messages are broadcasted periodically, and if a particular node fails to receive several such messages from a neighbor, a link breakage can be detected. Once a link breakage is detected, its routing table can be updated by deleting the entries in the routing table for that neighbor. This is how routing table is managed in the proposed system.

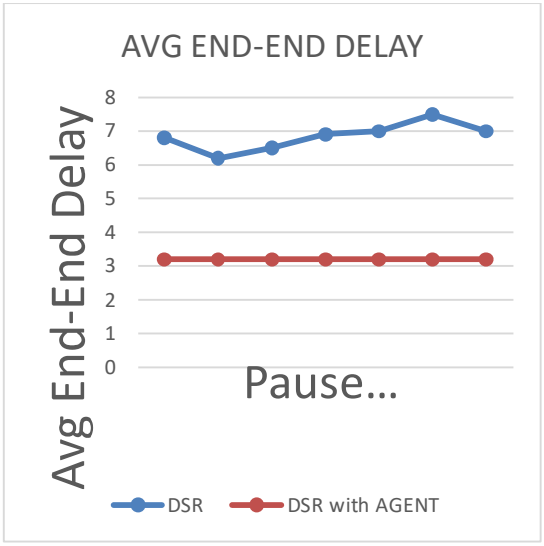
VIII.SIMULATION RESULTS

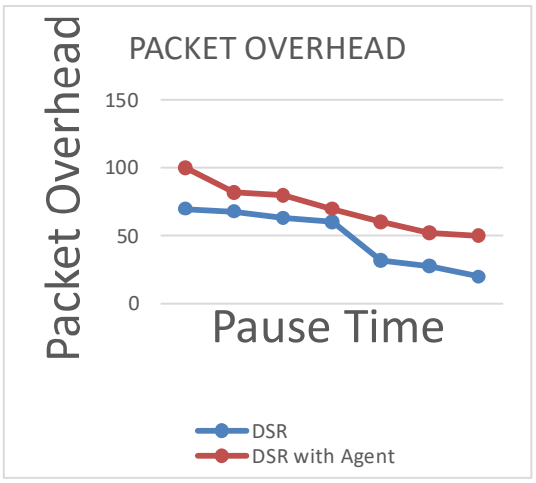
We simulate an Ad hoc network, the node behavior and mobility using ns-2.We found that the trustworthiness of the node is maintained. Also, we found that the packet delivery ratio ,average end-to-end delay, packet overhead .

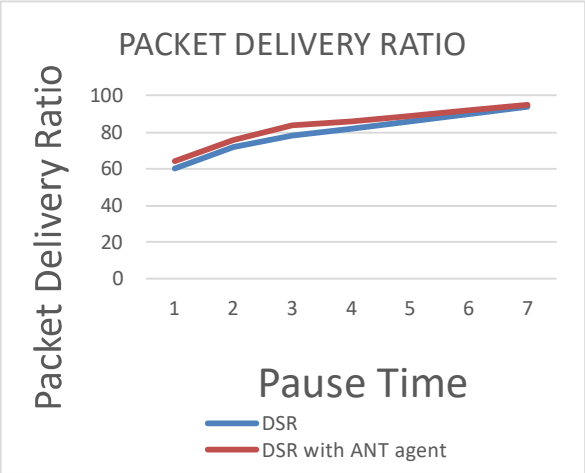
1. Packet delivery ratio: Representing the ratio of data packets received at its destination successfully.

2. Average end-to-end delay: Representing the average time taken for data packet before it was received at its destination.

3. Packet overhead: Representing the number of transmissions of control routing packets.







IX.CONCLUSION

In this paper we used the TSM mechanism to find the trustworthiness of the nodes.We discussed about the DSR routing protocol. The usage of Dynamic Source Routing using ANT agents to TSM mechanism helps to ensure confidentiality to a greater extent. We implemented asymmetric key techniques which helps to secure the data more precisely. Here, using the above said approach, we are able to improve the packet delivery ratio, reduce average end to end delay, and decrease packet overhead to a greater extent. Hence the overall performance of the system can be increased as well.

IX.REFERENCES

[1] Hong X,Kong J,Geria M.Mobility Changes anoaymity:new passive threats in mobile Ad hoc networks, Wireless Communications & Mobile Computing, 2006-05-01, 1530-8669.

[2]AhmedAbosamra,Mohamed Hashem,Gamal Darwish.Securing DSR with mobile agent in wireless Ad hoc networks,Egyptian Informatics Journal ,27 march 2011.

[3] Djatmiko, M., Boreli, R., Seneviratne, A., & Ries, S. (2011). Trust-based content distribution for mobile ad hoc networks ,modeling, analysis simulation of computer and telecommunication systems.

[4]Drs.Baruch Awerbuch & Amitabh Mishra , Department of computer science,Johns Hopkins University,Stamos J, Gifford D. Remote evaluation

[5]FuLiang Guo,Bin Zeng, LiangZhong Cui.A Distributed Network management Framework Based On Mobile Agents

[6]Atul Mishra, A.K.Sharma,Application of Mobile Agent in Distributed Network Management,Researchgate, DOI: 10.1109,2012.

[7]XU Xiao-long, XIONG Jing-Yi,CHENG Chun-Ling, The Model and the Security Mechanism of the Information Retrieval System based on Mobile Multi-Agent.

[8] [2]David B. Johnson, David A. Maltz. Dynamic Source Routing in Ad hoc Wireless Network,Researchgate,1999.

[9] Mentari Djatmiko, Roksana Boreli , Aruna Seneviratne, Sebastian Ries, ,Resources-aware trusted node selection for content distribution in mobile ad hoc networks.

[10] S. Menaka, Dr. M.K. Jayanthi, Intelligent Routing using Ant Algorithms for Wireless Ad Hoc Networks, Computer Network and Information Security, 10.5815,August 2013.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Spotting of Question Twin which has Semantically Similar

Vibha Lakshmi S Gopi
Department of Computer Science and Engineering, Faculty of Science and Technology,The ICFAI Foundation for Higher Education,Hyderabad – 501203
vibha.lakshmi15@ifheindia.org

S.Vairachilai
Department of Computer Science and Engineering, Faculty of Science and Technology,The ICFAI Foundation for Higher Education,Hyderabad – 501203
vairachilai@gmail.com

Brahma Naidu Kakarla
Department of Computer Science and Engineering, Faculty of Science and Technology,The ICFAI Foundation for Higher Education,Hyderabad – 501203
brahmanaidu@ifheindia.org

***Abstract*—Everyday more than 35 million people visit online forums such as Quora, stackoverflow, yahoo answers, Fluther, stackExchange and TED. In this forum, people may post number of questions that may contain identical words. The multiple questions with the same objective can cause answer seekers to spend more time to find best answer to their question and also make critic feel that they have to answer multiple versions of same objective question. These websites encourage canonical questions because they provide a better experience to active seekers and writers and offer more value to both groups in the long term. To prevent duplicate questions from existing, there is a need to develop the machine learning and natural language processing systems to identify the question automatically which has same intent have been asked multiple times. In this paper, identifying the question pair is the key focus using Siamese Long Short Term Memory Algorithm using pre-trained model to embed the words in the question.**

Keywords— Machine Learning, Natural Language Processing, Long Short Term Memory Algorithm

I. INTRODUCTION

There is a notable increase of substitute to classical web search to community based question answering. The questions are asked form various categories such as technical, health, science, life, history, mathematics, travel and many more. Community based question answering sites have collections of huge amount of questions and its related crowd sourced answers with time. The websites like StackExchange, StackOverflow is a sub domain that is exclusively meant for programming community-based answering. In case of Quora, various questions are posted by diverse users which may have similar meaning it is necessary to identify duplicate questions from the collection of questions. Random forest model is used to identify duplicate questions to determine whether the questions are duplicate or not. The best possible way is to find out the semantic meaning of the question.

Dasha Bogdanova et al. (2015) proposed a solution for detecting semantically equivalent questions (DSEQ) [1] in online user forums based on conventional neural network (CNN). This method handles the semantically equivalent questions problem as follows: First, the two questions are converted into tokenized text strings which are used as an input for the network. Second, the conventional neural network architecture transforms words into word embedding's (real valued feature vector) and builds two distributed vector representations for two given questions. Third, semantically similar question pair is identified based on the threshold value. Finally, similarity scores of vector representations are calculated by CNN.

Shashi Shankar et al. (2017) proposed solution to identify quora question pairs having same intent [2] by applying natural language processing techniques. This method handles the text similarity classification problem as follows: First, the dataset is divided into training (70%) and test set (30%). Second, features are extracted from the data such as finding length of question, difference in lengths, number of unique characters, number of words, longest common sub-string, Longest common subsequence and number of common words between two questions. Third, semantic similarity measures are applied on question pair and the feature vector are fed into the support vector machine (SVM). Finally, SVM classifier able to classify the unseen date such as duplicate or not.

Lei Guo et al. proposed solution to detect the duplicate quora questions [3] based on K- Nearest Neighbor (KNN,) SVM, Random Forest Classification algorithm, Decision Tree and Adaboost. The KNN gives the accuracy rate of 82.8%, hence KNN is chosen from other classifications methods for further predictions. They handled the text similarity classification problem as follows: First, vectorize questions in order to extract features, train and predict question vectors using machine learning techniques on features previously built. Second, implement the Word2Vec model on input to get vector representations of each word. Finally, it finds out the semantic similarity of words by determining cosine of two-word vectors.

The rest of the paper as follows: The computational details are discussed in section 2.

II. COMPUTATIONAL DETAILS

A. Word2Vec Model

The Word2vec is a neural network model that process the text in the dataset and generate word embeddings. The dataset consists of pair of questions. If a text corpus is given as input, it returns a set of vector representations of each word in the question. These vectors are used to obtain semantic similarity of words by calculating the cosine distance of two word vectors. The Quora dataset is used to train the Word2Vec model. In modeling, vectorization of words play a major role. The Word2vec is useful when analyzing the meaning of single word but semantic of sentence is not obtained directly by the sum of words vector. The parts of speech and order of each word together effect semantics.

Step 1: Vectorize the words in the question set based on Google's Word2Vec model.

Step 2: Take the mean of all words vectors to construct a vector for the question from its words vector.

Step 3: Compare the vector representations are identical or not.

Step 4: Identify the semantically equivalent questions.

In this method the effect of word sequences in the question are not considered. In recurrent neural network (RNN), to identify the vectorized questions embedding the Long Short Term Memory (LSTM) is used. Assign distinct id for each word in the question based on the frequency of its occurrence, the ids with the same frequency will be randomly placed. The embedding weight is based on Word2vec model. Finally fit the input to the model and evaluate the LSTM model.

B. Dataset description

The dataset is downloaded from Kaggle [4] website, Quora Question Pairs Competition. The dataset consists the following valid question pair data:

- Testing dataset – 1048576
- Training dataset – 404291

In training dataset, each row has column labels such as id, qid1, qid2, question1, question2 and is_duplicate. The test dataset contains same column labels as training dataset excluding "is_duplicate" label. In training and testing dataset, the iindependent variables are id (unique id of question pair), qid1 and qid2 (unique ids of each question). In training dataset, the independent variables are question1, question2. In training and testing dataset, the dependent variable is Is_duplicate. It is the target variable that is set to 1 if question1 and question2 are semantically same, otherwise 0. The question1 and question2 text will be cleaned initially before training of machine learning models approaches such as stemming and stop words.

C. Exploratory Data Analysis

The sample duplicate and non-duplicate questions are shown in TABLE 1. The number of occurrence of duplicate and non-duplicate questions in the dataset is shown in Fig 1. In Fig 1, the number of duplicate questions in the training dataset is 149263, the number of non-duplicates in the training dataset is 255027, the percentage of duplicates in training dataset is 36.92% and percentage of non-duplicates in training dataset is 63.08%. The Fig 2 indicates number of occurrences of duplicate and non-duplicate questions in testing dataset. The percentage of duplicates in testing dataset is 0.0% and number of duplicates is 0. The number of non-duplicates in testing dataset is 3563475 and its percentage is 100%.

TABLE I. Sample duplicate and non-duplicate questions

Duplicate question pair	**Non-duplicate question pair**
Pair 1: How can I improve my English language? How can I increase my English fluency? **Pair 2:** How can I be a good geologist? What should I do to be a great geologist? **Pair 3:** Which are the best ways to lose weight? How do I reduce my weight?	**Pair 1:** How much margin can I get by taking amul franchise? How automata theory applicable in text processing? **Pair 2:** If average velocity is zero, then is average acceleration zero? If velocity is maximum why does the acceleration have to be zero? **Pair 3:** What is the next business trend? What is the next business trend in Asia?

Fig. 1. Occurrences of duplicate to non-duplicate questions in the training dataset

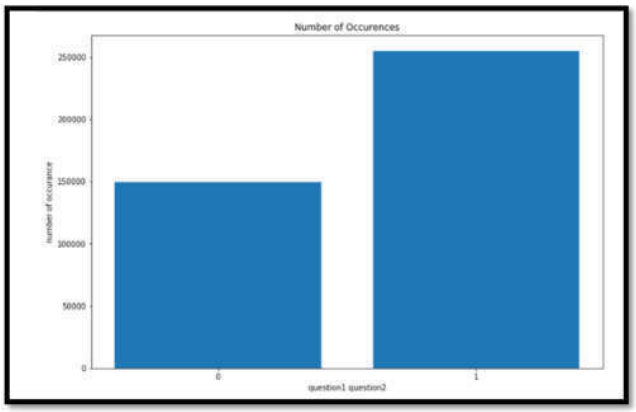
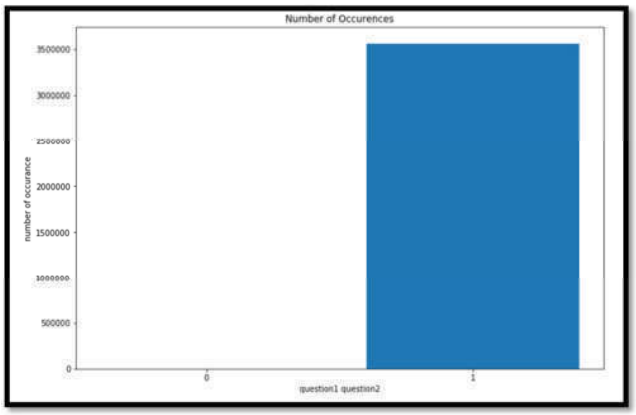


Fig. 2. Occurrences of duplicate and non-duplicate questions in the testing dataset



D. String based measures

In string based measure, the length of the questions, the absolute difference in the length of the two questions and the number of unique characters in all the questions are analyzed. The length of questions against the number of count of length shown in Fig 3.

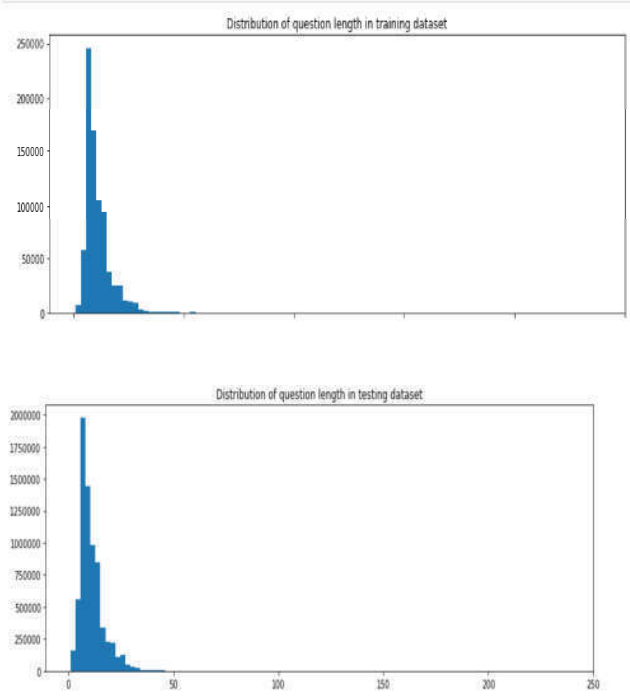


Fig. 3. Length of questions against the number of counts of length

III. METHODOLOGY

A. Data preprocessing

It is essential that without compromising on the potential predictive power of empty inputs, the null values are replaced with empty strings in order to ensure that null values are managed in proper manner. The dataset used in this analysis is pre-processed dataset there is no need to stress on data pre-processing.

B. LSTM model

The Siamese Manhatten Long Short Term Memory (LSTM) is used for modeling since gives a straight forward approach. The keras library has ease to train neural nets; the Siamese maLSTM is constructed by word embedding which is pre-trained by adopting Google corpus which has three million English words. The pre-trained model means that it is not necessary to train my own word embedding which results in long time for a model powered by CPU. The initial step is to define architecture of Siamese maLSTM model. This classifier does not provide explainable percept as a result.

The LSTM layer prevents over fitting the training date. In this model, there are two similar LSTM network. It is passed vector representations of sentences and output semantic description of the sentences as hidden state encoding to give similarity score. The questions are processed to process for the model by oobtaining the word matrix for the word embedding's. The architecture and the flow chart of LSTM model is shown in Fig.4 & Fig.5.

The process of obtaining word embedding's is accomplished by following steps:

Step 1: The splitting percentage is 80-20 of training dataset (train.csv), utilizing 20% as validation set to ensure better generalization.

Step 2: Cleansing of dataset by importing Natural Language [5] ToolKit (NLTK) stopwords along with some regex formatting to make sure that words are consistent.

Step 3: Iterate through every question of both training and testing datasets.

Step 4: The word vectors have to be converted into a numerical representation before training.

Step 5: Hold the numerical representation of each word and dictionary record of each word.

Step 6: Fill the matrix with each word's dimensional representation.

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 213)	0	
input_2 (InputLayer)	(None, 213)	0	
embedding_1 (Embedding)	(None, 213, 300)	36427800	input_1[0][0] input_2[0][0]
lstm_1 (LSTM)	(None, 50)	70200	embedding_1[0][0] embedding_1[1][0]
merge_1 (Merge)	(None, 1)	0	lstm_1[0][0] lstm_1[1][0]

Total params: 36,498,000
Trainable params: 70,200
Non-trainable params: 36,427,800

Fig. 4. Architecture of Siamese LSTM model

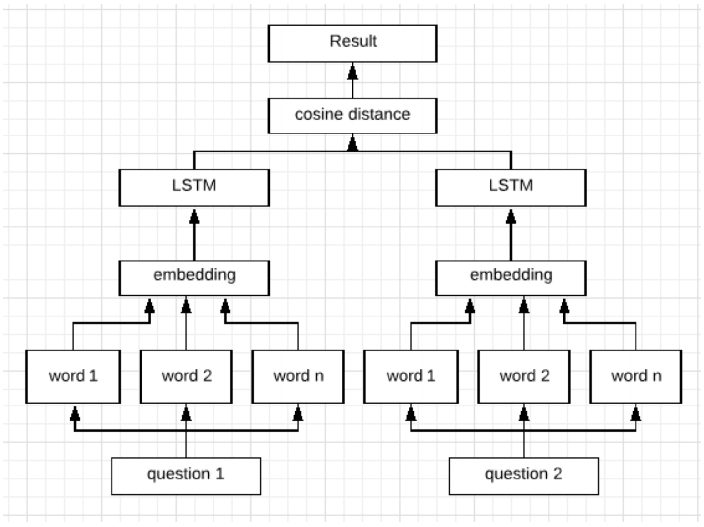


Fig. 5. Flow Chart of LSTM model

IV. RESULT

The number of epochs chosen would allow validation loss to stabilize by locating some better local optima or not moving back to shallower local optima. The Adam optimizer is used by clipping norm option is adopted as a counter measure against exploding gradient. The final model's parameters are: 50 hidden layer nodes, batch size 128, 15epochs, and Adam optimizer – clipping norm at value of 0.9. The weighted Log-Loss error is shown in the TABLE II. The model accuracy plots is shown in Fig.6 and the model

loss plots is shown in Fig.7. In Fig.6, X-axis represents number of epoch and Y-axis represents model accuracy value. In Fig.7, X-axis represents number of epoch and Y-axis represents model loss value.

TABLE II. Comparing metrics and results

Metric Model	Weighted Log-Loss Error
Siamese MaLSTM pre-trained word embedding	0.290278734785

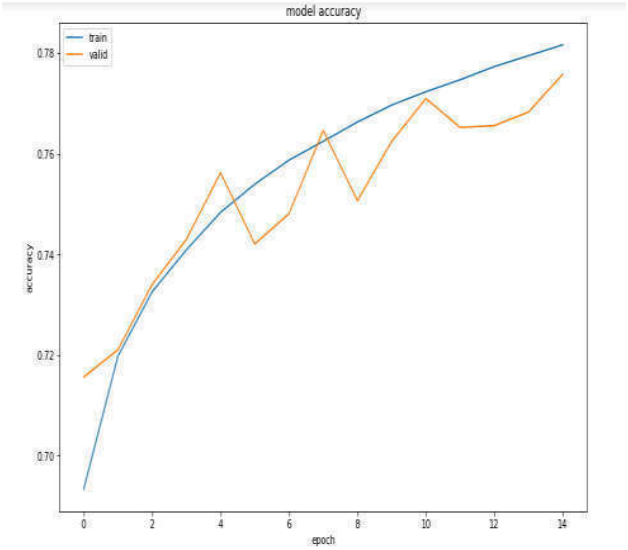


Fig. 6. Model accuracy plots for training and validation

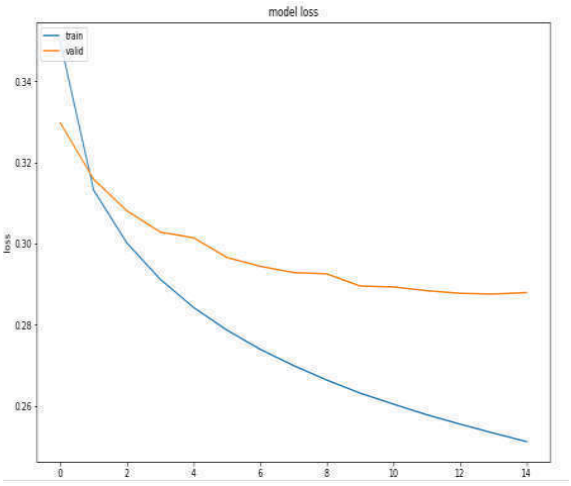


Fig. 7. Model loss plots for training and validation sets

Neural networks are a challenging area to deal with. Siamese network consists of similar networks, each network taking inputs. This network used for face recognition and image recognition which is a part of neural networks. Its main aim is to find similarity between the inputs.

In model accuracy, it is always better to see the validation accuracy being lower than the training accuracy. It relates to the notion of inability to generalize well enough and over fitting. The constant spine and troughs after 4 epochs means that the optimizer moved out of a local minimum to some higher or lower local minima, unable to gain the global minima. When compared with model loss, it tells us loss values against the number of epochs. The validation loss gives a good fit as it is not too far off from the training loss. The training loss value continuously drops as the model may tend towards over fitting the data. The model output Scalar test loss is 0.287926488689 and the mmodel accuracy is 0.775792129408.

V. Conclusion and Future Work

Long Short Term Memory (LSTM) algorithm which is constructed by word embedding used to identify whether the question pair is identical or not. Initially, the each word in the question is vectorized on Google's Word2Vec model. A vector is formed for the question by finding out the mean of all word vectors then the vector representation of two questions are compared to see if there are semantically similar or not. This paper discusses how to identify if the question pair is identical or not [6], the next step is to retrieve single, individual and independent canonical question from huge data. After identifying canonical questions new questions may be posted in the forum, this will lead us to check the quality of the question whether to retain or replace the canonical question.

References

[1] Dasha Bogdanova, C'icero dos Santos, Luciano Barbosa and Bianca Zadrozny "Detecting Semantically Equivalent Questions in Online User Forums" Conference paper 2015, in press

[2] Shashi Shankar and Aniket Shenoy "Identifying Quora question pairs having the same intent", unpublished

[3] Lei Guo, Chong Li and Haiming Tian "Duplicate Quora Questions Detection".

[4] https://www.kaggle.com/c/quora-question-pairs

[5] Collobert, R, &Weston, J.(2008). A unified architecture for natural language processing. Preceedings of the 25th international conference on Machine Learning – ICML 08. doi: 10.1145/1390156.1390177

[6] Y.Zhang, D. Lo, X.Xia and J.Sun. Multi-Factor Duplicate Question Detection in Stack Overflow. Journal of Computer Science and Technology, 30(5):981 -997,2015.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

FAKE CURRENCY DETECTION BASED ON FLUORESCENCE IN HSV COLOR SPACE

Dr. D. Stalin Alex
Dpt of IT
GNIT, Hyderabad.
drstalinalex@gmail.com

Dr. P. Subramanian
Dpt of CSE
Sri Indu CET,Hyderabad.
69subbu@gmail.com

Mrs. R. Sravani
Dpt of CSE,
GNITC(Autonomous),Hyderabad.
sravanirachakonda@gmail.com

***Abstract*— Fake currency notes are so perfect nowadays that it is very tough to differentiate them from original currency notes. Due to technological advancements, it has become very easy for counterfeiter's to imitate all the characteristics of the original currency except the illumination or glow that occurs when the currency notes are induced to UV radiation. The proposed approach is based on this selective feature of color illumination which can be identified when the captured image is converted into HSV color space. Histogram equalization is done in HSV color space which collectively performs noise reduction, filtering and sharpening to enhance image quality for effective use in applications. The original currency will have high intensity color values compared to the fake currency that can be verified by using the thresholding technique. Thus, the proposed method is very simple, efficient and time saving.**

Keywords*— *HSV (Hue, Saturation and value); color space; histogram equalization; threshold; fake Currency;

I. INTRODUCTION

Improvisation of the economic system is very important in producing the financial growth and giving social harmony. Automatic dispensers capable of recognizing currency notes are extensively used in many automatic banking operations. The need for automatic currency recognition system has been increased in order to develop a simple and reliable technique. Speed and accuracy are the two major factors to be considered in currency recognition. The currency recognition mainly focuses to extract the visible and hidden marks on paper currency for efficient classification. Until now, there are many methods proposed for paper currency recognition. The simplest technique for fake currency detection is by using the size and features of the currency. However, this technique has limitations as currency notes are getting worn and torn with the passing of time. During circulation, the currency notes have chances of losing information or even damaged. Moreover the complex design in currency notes makes automatic currency recognition difficult to process. So it is very important to know how the characteristics are extracted from the currency image using efficient algorithms to improve the accuracy of fake currency detection. The proposed method is simple and efficient to meet the high speed requirements in practical applications. Digital image processing is an area characterized by the need for extensive experimental proposed solutions to the given problem. Image processing has become very useful in research and industrial applications. In Digital image processing the input is image and the output maybe an image or the attributes extracted from the image. The proposed method in this paper deals with the analysis of the color feature currency notes. The process of extracting the color from paper currency is a challenging task in image processing. The algorithm we apply here is very simple and works efficiently. The currency image can be acquired by using a camera or a scanner so that the color feature of currency can be processed. Now the captured image can be processed by applying the techniques like image pre-processing, HSV color space, histogram equalization and color thresholding.

II. METHODOLOGY

The Indian currency notes have a special feature called ultraviolet fluorescence. Embedded fluorescent fibers into the paper, or printed ultra-violet ink onto the paper, creates a form of optical illumination or glow which makes verification easy. By exposing the note to ultra-violet light, the ink or fibers fluoresce, reveals a colored pattern which is not visible under natural light. In the proposed system, the above specific feature of color illumination can be identified by converting the captured currency image into HSV (Hue, Saturation and Value) color space. Color histogram equalization is done in the HSV color space in order to get an enhanced and rich image for extracting the color information of the image very clearly. Finally, thresholding technique is used for extracting and comparing the color intensity of the given currency image with the color intensity of the original currency. Therefore, the fake currency can be easily detected automatically without the need of existing UV detectors. Hence, the proposed system enables automatic detection of fake currency.

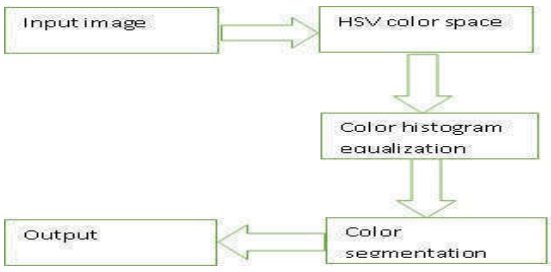


Figure 1 : System Architecture

ISBN 978-81-939386-1-4

A. *HSV color pattern*

In the proposed system, the properties of the HSV (Hue, Saturation and Value) color space of the currency note is analyzed based on the visual perception of the variation in Hue, Saturation and Intensity values of an currency image pixel. The pixel values are extracted by choosing the Intensity as the important property based on the Saturation value of a pixel. Therefore, a one-dimensional histogram is generated in the HSV space so that a smooth transition of color can be obtained. This enables us to match similar colors of the currency images. The Saturation threshold is used to determine the transition which is usually dependent on the Intensity. Human eye is less sensitive to saturation variation when compared to variation in Hue or Intensity. Therefore, we use the Saturation values to determine whether the Hue or the Intensity has more vision to human perception of the color of that pixel. Thus, HSV color space makes an approximation in the color of each pixel in the form of thresholding.

B. *Histogram Equalization*

For histogram generation in HSV color space each pixel contributes either its Hue or its Intensity value. The color histogram generated here, retains the color features of the currency notes without loss of information. The discrete values of Hue varies between 0 to 2π, the end points being red. Thus, the feature is a combination of two independent vectors. The color histograms generated for the currency image will have high values. This is because two colors that appear similar to human eye can have a small difference in their shade. The two histograms are compared using smoothing windows, so, that the color feature of the currency image is retained. The Hue-Intensity boundary of the image can be determined by using the above technique. Thus, by using HSV color space, the results generated have a better performance compared RGB color space.

C. *Color Segmentation & Thresholding*

Color segmentation is used to disintegrate an image into useful parts for analyzing and representing the currency image in terms of higher-level pixels by separating background and foreground objects. The HSV color space is prominently used because it can separate the intensity (luminance) from the color. For each pixel, either its hue or the intensity can be used as dominant feature based on its saturation. Finally, local thresholding method is applied to each pixel in the segmented currency image and the threshold T is obtained. Now, T is compared with the original currency threshold and if the value of T is less than the expected threshold, the currency is fake.

III. RESULTS AND DISCUSSION

Original Currency can be acquired by any device like cam scanner, scanner or capturing image by camera. Figure 2 shows the original image captured by normal mobile camera. Acquired image should not be in binary or grayscale image, it should be in color scale image. Binary or grayscale image gives the false result.



Figure 2 : Scanned Image of Original Currency

Acquired color image converts into HSV- fluorescence image for further processing. Figure 3 shows the HSE-Fluorescence view of original currency.

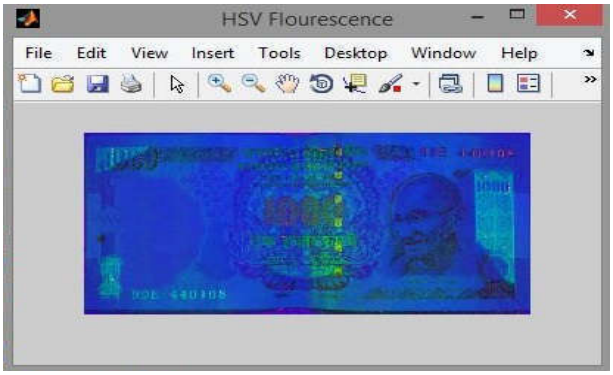


Figure 3: HSV Fluorescence view of original Currency

Fake Currency acquired by any acquisition device, Figure 4 shows the fake currency's original view.



Figure 4: Fake Currency

Now, the acquired fake currency converts into HSV-Fluorescence view. Figure 5 shows the fluorescence of fake currency.

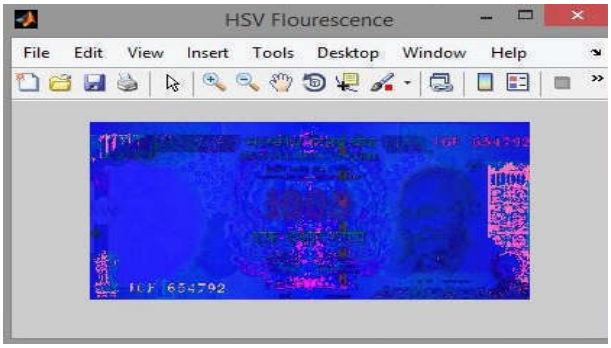


Figure 4: HSV Fluorescence view of Fake Currency

IV. CONCLUSION

We have proposed a simple and accurate technique which adopts HSV color model and thresholding technique. The system uses the interactive techniques such as histogram equalization and color thresholding in order to extract the color of the currency notes and detect fake currency. This technique can be equipped efficiently in automatic dispensers or cash deposit machines for counterfeit detection of Indian paper currency.

References

[1] Hardin, R. W., "Optical Tricks Designed to Foil Counterfeiters".OE Reports Number 191, International Society for Optical Engineering, November 1999.

[2] Murdoch, S. J., "Software Detection of Currency". University of Cambridge Computer Lab, 2004.

[3] Robert Fiete," Identifying falsified images can be straightforward if you know a few trick", from oemagazine , SPIE Newsroom,31 January 2005.

[4] L. Burke Files, "A manual for the Identification of Counterfeit Currency, Credit Cards, Traveler's Cheques and Bank Checks". Kumar Parasuraman, Member, IEEE and P.Vasantha Kumar, 2017.

[5] A.Ms.Trupti Pathrabe and B.Dr. N.G.Bawane, "Paper Currency Recognition System Using Characteristics Extraction and Negativity Correlated NN Ensemble",Int. Journal of Latest Trends in Computing, 2010.

[6] Angelo Frosini, Marco Gori, "A Neural Network-Based Model for Paper Currency Recognition and Verification", IEEE Transaction on NN Vol. 7 No.6, 2016.

[7] Dr.Kenji Yoshida, Mohammed Kamruzzaman, Faruq Ahmed Jewel, Raihan Ferdous Sajal, "Design and Implementation of a Machine Vision Based but Low Cost Stand Alone System for Real Time Counterfeit Bangladeshi Bank Notes Detection" IEEE 1-4244-1551-9/07/2007.

[8] Euisun Choi, Jongseok Lee and Joonhyun Yoon,"Feature Extraction for Bank Note Classification Using Wavelet Transform" @ISBN ISSN: 1051-4651, 0-7695-2521-0, IEEE, 2006.

[9] Ahmadi and S.Omatu," A Methodology to Evaluate and Improve Reliability in Paper Currency Neuro-Classifiers" Proceedings 2003 IEEE International Symposium on Computational intelligence in Robotics and Automation , Kobe, Japan,July 16-20,2003.

[10] S.Sural, G.Qian and S.Pramanik, "A Histogram with Perceptually Smooth Color Transition for Image Retrieval", Proc. Fourth Int. Conf. on CVPRIP, Durham, 2002.

[11] M.Swain and D.Ballard, "Color Indexing", Int. Journal of Computer Vision, vol.7, no. 1, pp. 11-32, 1991.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

EFFECTIVE APPROACHES FOR SECURING SENSITIVE DATA AND SUPPORTING MULTI DATA APPLICATIONS IN CLOUD FORENSICS ENVIRONMENT

[1]Sonya A
Computer science Engineering
BSA Crescent institute of science and technology,
Chennai-India.
sonyayasmin152@gmail.com

[2]Kavitha G
Information technology
BSA Crescent institute of science and technology,
Chennai-India.
gkavitha.78@crescent.education

Abstract*—Cloud computing became the most talked about innovative technology in Information and communication technology in current trends. The major concern of the different organizations now is the security of a user details or documents. In digital forensic investigations and log monitoring is presently an important and public concern. The impact is that patient's details doesn't work be actually processed and scattered nature. Most of the applications need to connect with several various data storage process depending on the type of data they have to maintain: existing data types, result report, statistical data from social networks etc. The importance of using provenance data is significant it provides the forensic investigators with data history in terms of patients, entities and activities involved in producing related patients health data reports. It is hard to locate the user's data where supposed to be stored. In this paper, propose a set of methods and encryption scheme is aiming at alleviating developers task for developing, deploying and migrating multiple data storage in real time tasks and also to enhance a secure provenance scheme based on advanced VM forensic digital examination analysis tools and block chain technology. It helps to manage and monitor the inner location of data, both where it comes from as it passes through the data provenance via hypervisor.

Keywords- Cloud Forensics, VM forensic digital examination analysis tools, Block Chain Technology and cryptography schemes.

I. INTRODUCTION

Cloud Computing offers so many benefits for real time applications and also various business organizations by providing very high scalable infrastructure resources, pay by use method and minimum cost on-demand computing [1]. The theme of this concept is to examine how the patients history can be identify, analyses and then maintain, it can be used in the cloud in order to trace the previous records of the patient's information reports has travelled in the cloud. But the main security threat of cloud computing has become an increasing anxiety. Malicious individuals can easily take authority power to hack the user's or patients sensitive details in cloud. Many of these attacks are different forms like unwanted messages (spam message), advertisements and unwanted websites. Cloud digital forensics tools can be defined as an applied science for the identification, preservation, analysis and presentation of data while preserving the reliability of the patient's information records and maintaining an exacting sequence of protection for the user's data. However, cloud forensics also includes investigating entire patient's file systems, execution flow and history. It is difficult to imagine the hospital management system can save all its patient private data and keep the security protection for that information. Once the patients get check up, need not to be check again and again. The patient's details can access directly through their cloud system. It leads to time consumption. It can access only by using OTP or Password in other system. Virtual cloud can be installing when the power is shutdown or no electricity.[2] Thus, the concept of this discussion is to defeat and reducing the real time issues in cloud and identifies an appropriate solution that can be useful in upcoming technology with optimum utilization resource to store the patients details with a secured manner. [3] It integrates numerous existent and new technologies such as hardware virtualization, mobile virtualization, desktop virtualization, etc and centralized computing. It provides endless 'unbelievable' sources to users as services across the Internet which helpful to get the details from patients. [4]A health care analytics reports are required to contain provenance details on the pathway the data took during various stages of processing of various tasks. Therefore, provide provenance to maintain the ownership and process history of patients data attributes in the data provenance store. We need to give protection and security for patient's general data and also private data. The requirements include integrity of documents, enforceability of the provenance record, and authentication of the user's identity. The main goal is to propose a practical secure provenance scheme with Hash Functions which can provide VM log monitoring for data forensics in cloud forensics. The data provenance records will provide digital investigators with a clear record of where the data has been travel in the cloud forensics. It also provides integrity on sensitive documents stored in a cloud. If someone like a hacker or unauthorized third parties trying to compromises data integrity or confidentiality of such a troublesome incident, the nature of work for forensics teams is to investigate then report and update the information. Our main concept is proficient exploitation in cloud forensics for getting desired aim of detection and prevention of forgery and criminals.

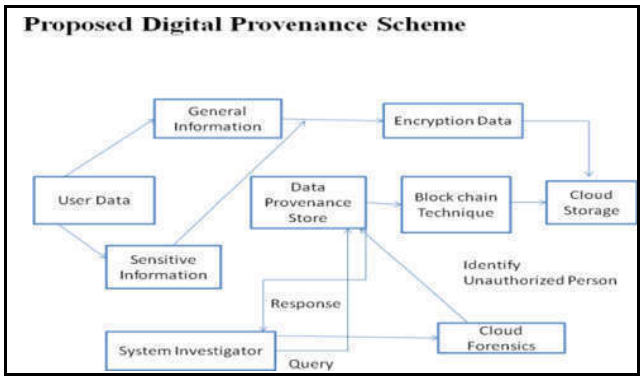


Fig 1. Diagrammatical way of proposed system

ISBN 978-81-939386-1-4

In the proposed scheme, user data are categorized into two ways like general information and sensitive information. Both general information and sensitive information are encrypted by hash functions stored in the cloud storage. It manages the data provenance store. It provides the support for storing the source data corresponding to the data provenance. The source data is stored in the Data Provenance Store which is also managed by a CSP. While both the User data and the system investigators are in the trusted environment. A user data has taken an input of source data, sends the corresponding data provenance to the Provenance Store. A system investigator may (1) send a request to the Provenance Store and as a result he/she (2) obtains the acknowledgement. Suppose if the system investigator identifies the unauthorized user trying to track the data. It takes the digital provenance using forensics process to find the unauthorized user. Finally Data forensics is stored in data provenance storage. When a collection of numerous patients' details is authorized to generate reports, if some issue arises in a modification, the concerned patient can be identified by some designated party using forensics investigation using block chain.

II. METHODOLOGY

1) Block Chain Technology

A block chain is an increasing chain of documents called blocks, which are chained using cryptography. Information held on a block chain exists as a distributed and continually subset in database. This technology database need not to be available in any single site, that means all the documents keeps are widely open and easily auditable and verifiable. Here there is no centralized report of this information exists for a hacker to crooked. Hosted by millions of computers concurrently, its information is accessible to anyone at everywhere by using internet. It protects our sensitive data, report and files, both on the nodes and in transmission, by using block chain technology and cryptography to encode the files. This system is visualized where every aspect of cloud storage such as transmission, generating or data storage is entered into the block chain. After this, which information should be analyzed, where it went, who accessed the information, and how that information was monitored can be checked by anyone who has authority power to generate to the block chain. Essentially, the block chain fixed the computing platform in time. It helps in giving complete identifiable, responsibility, and intelligibility for the cloud.All users and user's sensitive data in distributed cloud storage are interlinked network. It provides much reliable, faster, and less expensive than the existing systems. Thus, cloud storage enables users to store user's sensitive data in a secure manner. It is done by using block chain features such as transaction ledgers, hash functions, and cryptographic key encryption. Data centers are the hub of cloud storage capabilities for cloud giants like many open sources. But, these data centers are marketing with a high price tag for cloud developers, providers, and users. And also the impact of this higher cost of data failures and security breaches. The block chain is revolutionizing cloud storage by putting the user back in control over their data and devices. One of the reasons block chain is considered, because it is transparency to all with difficult manner. Anyone can access to the internet, it will check the chain and see for them whether everything up till now is justifiable. Another reason is because of the proof-of-work concept block chain work. The following are the three major benefits of distributed cloud storage.

2) Tamper-proof data

Backup and storage techs have proved that save user data has not been tampered with, when an effective and verifiable backup was introduced and created. Distributed cloud storage based on block chain technology save only hashes of its data blocks for users sensitive data. And the encoded and distributed hashes are important to verify these user data blocks.

3) Verifiability

Block chain is just not only save data in a distributed and encoded format, but also provides for linear chain in which every block contains a cryptographic hash functions. This links to the blocks and thereby, creates a decentralized transaction ledger.

4) No more middlemen

Many cloud techs, the biggest impact that block chains are likely to bring is to serve the product directly to the user. This is because a well-designed and publicly accessible block chain can replace .It mainly providing a trustworthy trading environment, guarding against fraud and handling, ensuring contract compliance, and financial transactions management system. Block chain's power does not lie in its difficult manner; its distribution across a chain of computers also makes block chain harder to attack. This technology is accessible by own self verifying sequential storage scheme.

III. MODULES DESCRIPITION

- Attributes definition/Master Secret key generation module
- Access structure Specification Module
- Private key generation module
- Transformation Key generation Module
- Retrieving Key Generation Module
- Final performance evaluation module

a) Attributes definition/master secret key

- An attribute is a one kind of data which access the properties of a field or tag in a database or a string of characters in a display.
- A key is considered as numeric that contains the output of an encryption algorithm when the condition is exchange from plain text to cipher text.
- Master secret key is a cryptographic key used in protected data communication.
- In this module, a master secret key is generated from the inputs provided by the user and the finalisation of attributes to be used for encryption.

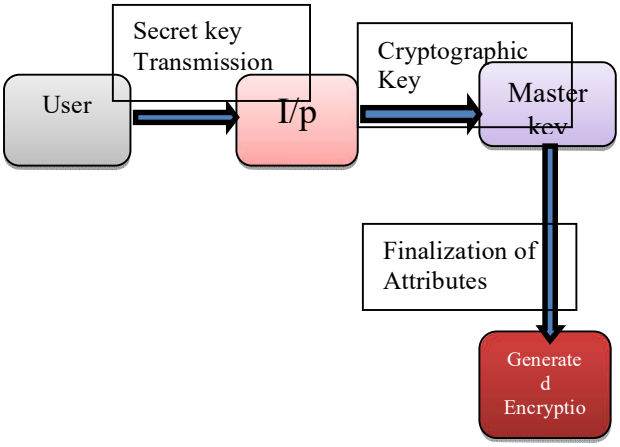


Fig 2. Attributes definition /Master secret key

b) Access structures specification module

- A detailed description of the design and materials used to make something is specification. Access structures are used to study the structure of security system. So where multiple workers need to work together to obtain a necessary source.
- Based on the access structure of the users the data will be encrypted for security purpose.
- So that, the algorithm will be defined dynamically for data encryption.
  The message to be passed is converted into cipher text.

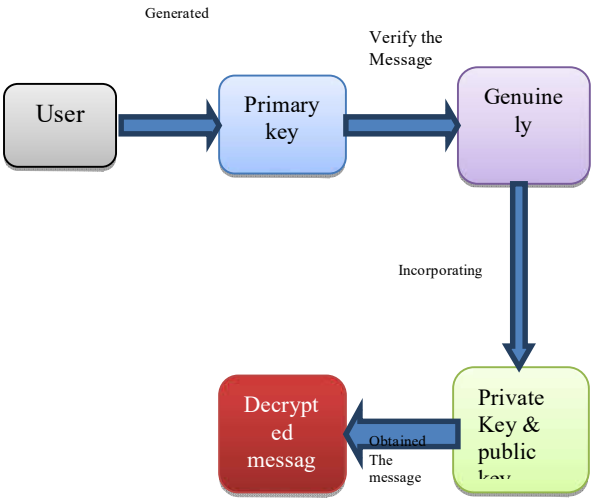


Fig 4. Private Key generation module

c) Transformation key generation module

➢ A Partialised algorithmic approach is used in this module so as to decrypt the cipher text into partially decrypted cipher text.

➢ The final data can be generated from the partially ciphered text.

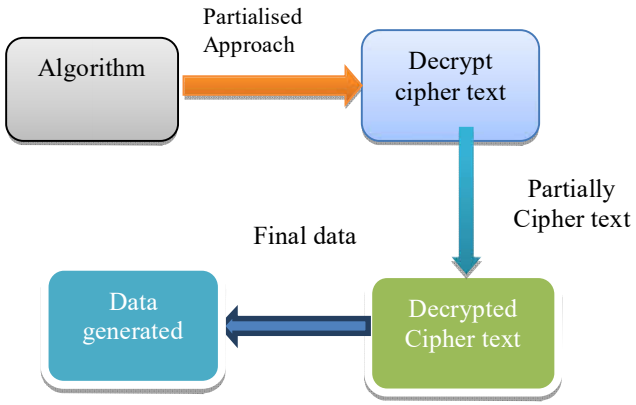


Fig 5. Transformation key generation

d) Retrieve key generation module

- Retrieve is defined as to find and bring or to bring back or restore, whether physically or mentally. The process of decoding data that has been encoded into a secured format.
- Retrieving key is a cryptographic key used to retrieve the partial decrypted information fully.
- In this module, Partial key can be decrypted further to receive the final message which is validated and compared with the decrypted one received in the third module.

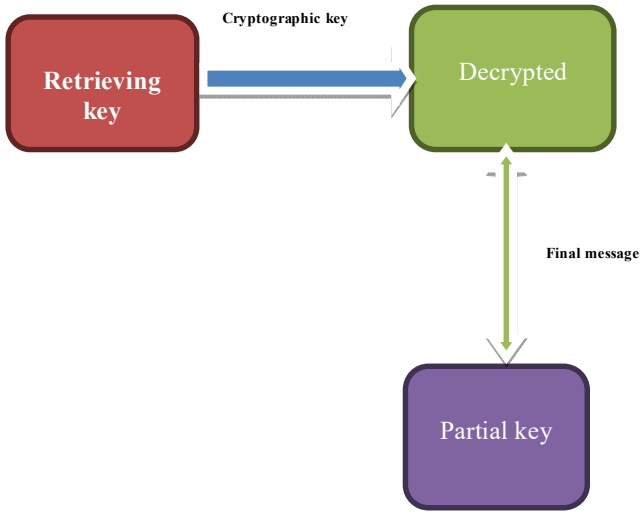


Fig 6.Retrieve key generation module

IV. Conclusion

Hence the objective considered thereby defining and verifying a patient document to check the patient details like attributes and master secret key using block chain technology depending on encryption and decryption techniques. It will achieve much private and fine-grained data access control, proficient and integrity the user data

with a secure manner. In this paper, propose a set of methods and encryption scheme is aiming at alleviating developers task for developing, deploying and migrating multiple data storage in real time tasks and also to enhance a secure provenance scheme based on advanced VM forensic digital examination analysis tools and block chain technology. It helps to manage and monitor the inner location of data, both where it comes from as it passes through the data provenance via hypervisor.

V. REFERENCES

[1] B. Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," in Proc. Public Key Cryptography, 2012, pp. 53–70.

[2] NuttapongAttrapadung, Javier Herranz, Fabien Laguillauimie, Benoit Libert, Elie de Panafieu, and Carla Rafols. Attribute-based encryption schemes with constant-size ciphertexts. Theor. Comput. Sci., 422:15-38, 2012.

[3] J Ayo Akinyele, Gary Belvin, Christina Garman, Matthew Pagano, Michael Rushanan, Paul Martin, Ian Miers, Matthew Green, and Avi Rubin. Charm: A tool for rapid cryptographic prototyping. Available from http://www.charm-crypto.com/, 2012.

[4] "State of blockchain q1 2016: Block chain funding overtakes bitcoin," 2016. [Online]. Available: http://www.coindesk.com/ state-of-blockchain-q1-2016/

[5] Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The block chain model of cryptography and privacy-preserving smart contracts," in Proceedings of IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2016, pp. 839–858.

[6] Kan Yang and Xiaohua Jia," An Efficient and Secure Dynamic Auditing Protocol for Data Storage in Cloud Computing" in the IEEE Transactions on Parallel and Distributed System, Vol. 24, No. 9, September -2013.

[7] S.Hemalatha and S.Alaudeen Basha,"Enabling for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud" Journal of Scientific and Research Publications,vol.3 issue.10,oct 2013.

[8] C.Celcia and T.Kavitha,"Privacy Preserving Heuristic Approach for Intermediate Data Sets in Cloud", Journal of Engineering Trends and Technology, vol.9, no.5, march 2014.

[9] A.Sonya, Dr. G.Kavitha, Efficient approaches for storing retrieval and replication of data using cloud system. International Journal of Pure and Applied Mathematics Volume 117 No. 20 2017, 1011-1020 ISSN: 1311-8080 ISSN: 1314-3395 Nov 17.

[10] Dineshkumar C, Subramanian M "Experimental Investigation of Onboard Driver Condition Monitoring System for Passenger Vehicles" International Journal of Mechanical Engineering and Technology (IJMET) Volume9, Issue6, June 2018, pp.01-09. Article ID: IJMET_06_07_001.

[11] A.Sonya Clustering of Crime Information by Using Mobile Application, International Journal of Engineering Research & Management Technology September-2018 ISSN: 2348-4039 Volume 5, Issue-5.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Cipher text policy for attribute-based encryption in cloud data

M. Raja,
Department of Computer Science and Engineering,
Kalasalingam Academy of Research and Eduction,
Krishnankoil, Tamil Nadu, India.

kingraaja@gmail.com

S.Dhanasekaran

Department of Computer Science and Engineering,
Kalasalingam Academy of Research and Eduction,
Krishnankoil, Tamil Nadu, India.

srividhans@gmail.com

Abstract— **With the confirmation of dispersed figuring, cell phones can spare/show signs of improvement singular actualities from everywhere at something component. Figure content material cloth fabric technique first-rate based encryption (CP-ABE) is probably a promising crypto logical approach for quality grained get to management of re-appropriated statistics inside the cloud. Be that as it can, a couple of dangers of key business enterprise disappointed the confirmation of its software. One impairment in crucial need of needs is that the important thing created potential downside. We exhibit that the front-quit gadgets of customers like practical phones quite a few the time have constrained warranty inclusion, so if individual keys are sincerely manage with the guide of them, clients peril key advent that isn't always in each case actually found yet typically existed in beyond exam. Moreover, huge purchaser deciphering overhead compels the realistic use of ABE. In those paintings, we will be inclined to prescribe an accommodating key company life-style in CP-ABE (CKM-CP-ABE). Our development recognizes passed on age, issue and functionality of character keys while no longer which includes any what's more repute quo. A fantastic grained and brief feature denial is proper key revive. The orchestrated beneficial component effectively settles no longer certainly key normal assertion downside at any charge conjointly key advent. In the duration in the center of, it lets in specifically lessen customer unscrambling overhead. A relationship with other agent CP-ABE designs shows that our association has in reality better execution as a long way as cloud-constructed re-appropriated realities taking an interest.**

Keywords—Cloud data sharing, CP-ABE, Key management, Security, efficiency.

1. INTRODUCTION

I. INTRODUCTION

With rate-plausibility overhauls in computational advancement and generous scale systems, providing data to others at closing at prolonged final finally ends up being correspondingly more noteworthy gainful. Also, automated possessions are certainly more without problem procured with the beneficial asset of techniques for scattered enlisting and confinement. Since cloud measurements sharing requires off-premises structure that or 3 affiliations collectively held, an all-inclusive way off putting away is through some strategies debilitating assure of statistics owners. Thus, maintaining the nicely-being of guy or woman, man or woman and touchy data set away within the cloud may be very crucial. The concurrent excitement of a significant massive sort of customers requires remarkable grained get the probability to control for realities sharing. Quality based actually honestly encryption (ABE) is a promising cryptographic excessive that offers a thrilling response for cozy and adaptable realities sharing. ABE has capability one-to-several assets, which deduces a single key can unscramble particular cipher texts or obvious keys can interpret the proportionate cipher textual content. There are kinds of ABE, known as cipher text system ABE (CP-ABE) and key method ABE (KP-ABE). For CP-ABE, the get entry to approach is mounted right at once right into a cipher text and the big set is added proper right away into a character key. For KP-ABE, the entry interest plan is conveyed appropriate at once right into a personal key and the mind blowing set is mounted valid straight away into a cipher textual content.

CP-ABE engages realities proprietors to depict their personal interesting way of improvement. Any person who wants to get information should within the route of the start installation together the direction device out of the normal set. As a give up consequence of this blessings, CP-ABE internal purpose turning into for the development of free, excellent grained get the risk to control for cloud realities sharing.

In this paper, we prompt a completely novel network located key enterprise industrial employer venture manner of existence in cipher textual content technique terrific essentially primarily based encryption (CKM-CP-ABE) seeking out to upgrade guarantee and execution of key organization place of business association in cloud certainties sharing shape. The well-known obligations are thick as seems for after:

An epic synergistic manner of lifestyles is established. With help of courting the first rate key deal with, a cloud server and a purchaser who has a bowed to get to records, scattered age, inconvenience and confine of character keys are made pleasure in of it. Along those follows, agreeable key challenge is guaranteed without together with any more noteworthy big snared installation, or, at the surrender of the day to stay away from on separated and beyond multi-grasp systems.

We acquaint trademark events with increment the non-public key animate preserve in thoughts. Tremendous assets collect thrillers allotted to each ability corporation enterprise assignment that consists of customers who charge a relative trademark. Through resuscitating outstanding social event key, a pleasant-grained and spark off potential denial is given.

3) we display that key escrow hassle and also key presentation is debilitating the riddle of private keys, or, all subjects considered in beyond studies. Seemed normally in expressions of beyond key manager customs for outcomes basically based totally completely realities sharing shape in cloud, our proposed way of lifestyles adequate has a dishonest to each hassle via its synergistic key undertaking.

At lengthy ultimate, we supply declaration of nicely-being to the proposed manner of life.

4) The communitarian stage empowers exceptionally decrease customer unscrambling overhead via the usage of using an unraveling server to execute thru strategies for way of the use of a long way the large majority of decoding at the same time as overlook no locating type of information to it.

II. PROPOSED SYSTEM

The model of our gadget is proven in Fig. 1. As depicted, five entities are involved in cloud information sharing:

Client: Client (CL) is a purchaser who hopes to get to records in allocated carport via the front-stop gadgets. With the usefulness instance of accessible cloud businesses, cell telephones are the more outstanding a chunk of the front-forestall deices. On the off chance that the CL's assets set satisfies at the front activity plan associated with figure literary substance, the CL may be jail to accomplish

plaintext. We anticipate that maximum cellular telephones are execution controlled, so CLs is possibly in danger of putting in with key advent [3].

Key Authority Client (CL) is a customer who hopes to get to records in allocated carport through the the front-forestall devices. With the usefulness example of handy cloud businesses, cell telephones are the more prominent a piece of front-quit deices. On the off threat that the CL's assets set satisfies a front recreation plan associated with parent literary substance, the CL may be prison to perform plaintext. We count on that maximum mobile telephones are execution managed, so CLs is probably in threat of putting in vicinity with key creation.

Cloud Server Cloud server (CS) is in charge of disbursed carport control. Each one of the insights to be shared is in the manipulate of the CS. We depend upon that any CS is semi-trusted.

Decryption Server : The decoding server (DS) has severe dealing with limits. It grasps and disconnects the most, but now not all undertaking of unscrambling. We depend upon that the DS is semi-depended on and the DS get the opportunity to channel is complex, for the motive that it's miles all right for CKM-CP-ABE to make certain records guarantee.

6) Data Owner Statistics owner (DO) is an authorized client in the device who has records to be exchanged. DOs include their personal taken into consideration one in every of a type one among a kind get right of segment to arrangements with the cause that in reality attractive CLs are yielded conform to get plaintext

7) assume that all substances associated with records sharing does not plot with every great to get to records illegally, for the maximum component the affiliation is probably blocked off and immaterial. In our model, attributes are installation thru the KA. Each and each yielded belonging are addressed with the guide of a assembly of self-assertive delivered materials coordinated into open parameters, or, in numerous phrases the KA as a hard and fast with a CS. Give Parma's a threat to be open parameters. At the hassle on the identical time as a DO hopes to charge facts, it encodes the records making utilization of Parma's dispatched to form the basic cipher textual content CT init and exchanges it to the KA. The KA re-encodes the essential cipher textual content to shape a conclusive cipher textual content CTulti, this is despatched to and determined away in a CS. , the number one trouble. According to the CL's large set 123 S manage life-style serves to the majority of the even as and speedy make three unique components of the non-public key, specifically, CPK1 , CPK2 and CPK3 , each one in every genuinely one in every of that is stored thru in truth one in the good deal approximately, CS or CL. When asked for facts observed away inside the cloud, the DS . gets CPK1 and CPK2 to opportunity CTulti to CT via its from CT Eventually, the CL isolates the plaintext CPK3.

For our proposed CKM-CP-ABE for cloud insights sharing, practically through the overall of each one of the three man or woman key fragments can plaintext be removed from a conclusive cipher text? It shows a CL requires facilitated endeavor with the KA and CS for unscramble cipher textual content.

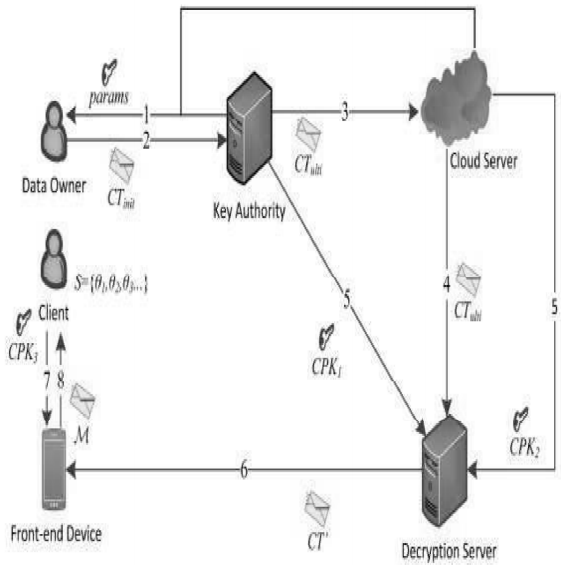


Fig. 1. The model of CKM-CP-ABE for cloud data sharing.

III. KEY MANAGEMENT ALGORITHM: -

Our CKM-CP-ABE scheme consists of six algorithms whose details observe.

A. Setup:-

Give very well a chance to be a wellness parameter.G1 and G2 are multiplicative cyclic businesses with unbalanced request p. Give g a hazard to be a generator of G1.

3) CS Setup:

The CS chooses a random exponent then, it computes a grasp secret/public parameter pair, that is:

$$(MK_{cs} = g^{a},\ PP_{cs} = e(g,g)^{a})$$

Let prams =(PP,PPKA,PPCS) be open parameters yield by method for manner of the setup set of commands. As laid out in reference [10], PP is hypothetically decided on with the useful precious asset of a nicely known as legitimate with launcher earlier than the advent of the plan. For our CKM-CP-ABE, we keep the bear in thoughts setup set of controls saved taking walks via strategies for utilizing a get as actual with launcher to keep hypothetical rightness. In useful acknowledgment, for the reason that does no longer consolidate any touchy facts, it is not critical to don't forget its period for privateness. Be that as it may, worldwide consistency need to be assured via manner of the association of the KA to create and speak PP.

B. Key Generation:

The key innovation set of pointers, it's far run through way of the KA, takes as input the general populace parameter PP and a CL's ability set S. At that point, the KA choices an strange instance at that aspect, the vital detail innovation set of tenets yields the primer key as pursues:

$$PK_{out} = (g^{\tau}, \forall x \in S : h_x^{\tau})$$

The initial key is kept by the KA for subsequent private key updates.

C. Encryption:-

The encryption figuring, or, in out of the regular expressions the DO, takes as records the general masses parameters params =(PP,PPKA,PPCS) a get admission to border and plaintext. At that thing, the DO produces a self-assertive vector and an immediately backbone chiller sharing affiliation associated with A. At that phase, the DO yields the hidden cipher textual content:

$$CT_{out} = ((M, \rho), C = M.e(g,g)^{ax}$$
$$C_{\tau} = g^{\tau}, \forall \theta_t \in A: C_{\tau} = g^{x}.h_{\rho(t)}^{-x}.$$

D. Re-encryption:-

The re-encryption figuring, or, in one-of-a-kind states the CS, takes as facts the general loads parameters params =(PP,PPKA,PPCS) the crucial cipher textual content and the development of highlight social events. We include the top notch combination basically situated as a famous rule depend of to re-scramble the essential cipher text.

A conclusive cipher textual content is discovered away in the cloud. At the factor even as the front inquiry is gotten, the CS will dispatch a healing. The goal of the re-encryption figuring is spotting terrific capacity repudiation.

E. Private Key Update:-

The man or woman key invigorate computation, or, in considered one of a type phrases improvement of CKM-CP-ABE, takes as insights the parameters params =(PP,PPKA,PPCS) and the essential key.

For this figuring, the tool key control way of life is carried out to make and scatter three specific man or woman key components. The manner of lifestyles incorporates of sub-traditions.

F. Decryption:-

The unscrambling rely number takes as records CL's brilliant set S, a conclusive cipher textual content and all non-open key elements On getting a decoding inquiry from a CL the KA and CS shipping their private key sections to the DS. At that element, the CS allows the DS to recoup the consultation key of o.K. U and an authoritative cipher text from circulated stockpiling.

Conclusion:

Cipher textual content technique property based totally completely encoding may be a promising mechanical understanding manner to see top notch grained get to control in comfortable administered carport. In the midst of this paper, we've got were given a penchant to underwrite a surely particular supportive key corporation culture to boost every coverage and productivity of key corporation in cipher text method fine especially primarily based totally cryptography for cloud actualities sharing form. Scattered key age, difficulty and limit of character keys rectangular diploma completed. however, no longer comprehensive of any additional giant system. We may be slanted to familiarize exceptional social activities with make a non-open key invigorate rule for excellent grained and start out feature disavowal. The ordinary charming gadget in reality addresses now not most effective key device drawback anyways similarly a in addition lousy downside mulled over key advent that past research barely tremendous. In the meantime, it progresses customers' purchaser enjoy given that exclusively a bit percent of responsibility is taken by the use of them for riddle developing. Thusly, the everyday situation plays better in cloud facts sharing form serving sizeable execution managed the front-stop gadgets with congruity each coverage or skill capability.

References:-

[1] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in Proc. EuroCrypt, 2005, pp. 457-473.

[2] J. Bethencourt, A. Sahai, and B.Waters, "Ciphertext-policy attribute-based encryption, in Proc. IEEE Symp. Secur. Privacy, 2007, pp. 321-334.

[3] N. Attrapadung and H. Imai, "Conjunctive broadcast and attribute-based encryption," in Proc. Int. Conf. Pairing-Based Cryptography, 2009, pp. 248-265.

[4] B. Waters, "Ciphertext-policy attribute-based encryption: an expressive, efficient, and provably secure realization," in Proc. Public Key Cryptography, 2011, pp. 53-70.

[5] M. Green, S. Hohnberger, and B. Waters, "Outsourcing the decryption of ABE ciphertext," in Proc. USENIX Secur. Symp., 2011, pp. 34.

[6] J. Lai, R. H. Deng, C. Guan, and J. Weng, "Attribute-based encryption with verifiable outsourced decryption," IEEE Trans. Inf. Forens. Security, vol. 8, no. 8, pp. 1343-1354, 2013.

[7] S. Lin, R. Zhang, H. Ma, and M. Wang, "Revisiting attribute-based encryption with verifiable outsourced decryption," IEEE Trans. Inf. Forens. Security, vol. 10, no. 10, pp. 2119-2130, 2015.

[8] M. Chase, and S. S. M. Chow, "Improving privacy and security in multi-authority attribute-based encryption," in Proc. ACM CCS, 2009, 121-130.

[9] G. Zhang, L. Liu, and Y. Liu, "An attribute-based encryption scheme secure against malicious KGC," in Proc.TRUSTCOM, 2012, pp. 1376-1380.

[10] J. Hur, "Improving security and efficiency in attribute-based data sharing," IEEE Trans. Knowl. Data. Eng., vol. 25, no. 10, pp. 2271-2282, 2013.

[11] P. P. Chandar, D. Mutkurman, and M. Rathinrai, "Hierarchical attribute based proxy reencryption access control in cloud computing," in Proc. ICCPCT, 2014, pp. 1565-1570.

[12] X. A. Wang, J. Ma, and F. Xhafa, "Outsourcing decryption of attribute based encryption with energy efficiency," in Proc. 3PGCIC, 2015, pp. 444-448.

[13] L. Cheung, and C. Newport, "Provably secure ciphertext policy ABE," in Proc. ACM CCS, 2007, pp. 456-465.

[14] J. Hur, and D. K. Noh, "Attribute-based access control with efficient revocation in data outsourcing systems," IEEE Trans. Parallel Distrib. Syst., vol. 22, no. 7, pp. 1214-1221, 2011.

[15] M. Pirretti, P. Traynor, P. McDaniel, and B. Waters, "Secure attribute-based systems," in Proc. ACM CCS, 2006, pp. 99-112.

[16] A. Boldyreva, V. Goyal, and V. Kumar, "Identity-based encryption with efficient revocation," in Proc. ACM CCS, 2008, pp. 417-426.

[17] A. Xiong, C. Xu, and Q. Gan, "A CP-ABE scheme with system attributes revocation in cloud storage," in Proc. ICCWAMIP, 2014, pp. 331-335.

[18] Q. Wu, "A generic construction of ciphertext-policy attribute-based encryption supporting attribute revocation," China Commun., vol. 11, no. 13, pp. 93-100, 2014.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A NOVEL PATH INFERENCE APPROACH TO RECONSTRUCTING THE PACKET ROUTING PATHS IN DYNAMIC AND LARGE SCALE NETWORKS.

Byreddy Madhavi
PG Student
Department of CSE
Malla Reddy College of Engineering
Hyderabad, India
Saimadhavireddy11@gmail.com

P. John Paul
Professor
Department of ECE
Malla Reddy College of Engineering
Hyderabad, India
principal@mrce.in

Abstract: **Recent wireless sensor networks (WSNs) are becoming increasingly complex with the growing network scale and the dynamic nature of wireless communications. Many measurement and diagnostic approaches depend on per-packet routing paths for accurate and fine-grained analysis of the complex network behaviors. In this paper, we propose I PATH, a novel path inference approach to reconstructing the per-packet routing paths in dynamic and large-scale networks. The basic idea of I PATH is to exploit high path similarity to iteratively infer long paths from short ones. I PATH starts with an initial known set of paths and performs path inference iteratively. I PATH includes a novel design of a lightweight hash function for verification of the inferred paths. In order to further improve the inference capability as well as the execution efficiency, I PATH includes a fast bootstrapping algorithm to reconstruct the initial set of paths. We also implement I PATH and evaluate its performance using traces from large-scale WSN deployments as well as extensive simulations. Results show that I PATH achieves much higher reconstruction ratios under different network settings compared to other state-of-the-art approaches.**

In this paper, we propose I PATH, a novel path inference approach to reconstructing the routing path for each received packet. I PATH exploits the path similarity and uses the iterative boosting algorithm to reconstruct the routing path effectively. Furthermore, the fast bootstrapping algorithm provides an initial set of paths for the iterative algorithm. We formally analyze the reconstruction performance of I PATH as well as two related approaches. The analysis results show that I PATH achieves higher reconstruction ratio when the network setting varies.

I.INTRODUCTION

WIRELESS sensor networks (WSNs) can be applied in many application scenarios, e.g., structural protection [1], ecosystem management [2], and urban CO monitoring [3]. In a typical WSN, a number of self-organized sensor nodes report the sensing data periodically to a central sink via multi hop wireless. Recent years have witnessed a rapid growth of sensor network scale. Some sensor networks include hundreds even thousands of sensor nodes [2], [3]. These networks often employ dynamic routing protocols [4]– [6] to achieve fast adaptation to the dynamic wireless channel conditions. The growing network scale and the dynamic nature of wireless channel make WSNs become increasingly complex and hard to manage. M Reconstructing the routing path of each received packet at the sink side is an effective way to understand the network's complex internal behaviors [7], [8]. With the routing path of each packet, many measurement and network manager can easily find out the nodes with a lot of packets diagnostic approaches [9]–[13] are able to conduct effective management and protocol optimizations for deployed WSNs consisting of a large number of unattended sensor nodes. For example, PAD [10] depends on the routing path information to build a Bayesian network for inferring the root causes of abnormal phenomena. Path information is also important for a network manager to effectively manage a sensor network. For example, given the per-packet path information, a forwarded by them, i.e., network hop spots. Then, the manager can take actions to deal with that problem, such as deploying more nodes to that area and modifying the routing layer protocols. Furthermore, per-packet path information is essential to monitor the fine-grained per-link metrics. For example, most existing delay and loss measurement approaches [9], [14] assume that the routing topology is given as *a priori*. The time-varying routing topology can be effectively obtained by per-packet routing path, significantly improving the values of existing WSN delay and loss tomography approaches. A straightforward approach is to attach the entire routing path in each packet. The problem of this approach is that its message overhead can be large for packets with long routing paths. Considering the limited communication resources of WSNs, this approach is usually not desirable in practice. In this paper, we propose I Path, a novel path inference approach to reconstruct routing paths at the sink side. Based on a real-world complex urban sensing network with all node generating local packets, we find a key observation: It is highly probable that a packet from node and *one of* the packets from 's parent will follow the same path starting from 's parent toward the sink. We refer to this observation as *high path similarity*. Fig. 1 shows a simple example where S is the sink node. Denotes a packet from A, and denotes packets from B (A's parent). *High path similarity* states that it is highly probable that will follow the same path (i.e., , which means the subpath by removing node A from) as one of B's packet, say , i.e., . The basic idea of I PATH is to exploit high path similarity to iteratively infer long paths from short ones. I PATH starts with a known set of paths (e.g., the one-hop paths are already known) and performs path inference iteratively. During each iteration, it tries to infer paths one hop longer until no paths can be inferred. In order to ensure correct inference, I PATH needs to verify whether a short path can be used for inferring a long path. For this purpose, I PATH includes a novel design of a lightweight hash function. Each data packet attaches a hash value that is updated hop by hop. This *recorded hash value* is compared against the *calculated hash value* of an inferred path. If these two values match, the path is correctly inferred with a very high probability. In order to further improve the inference capability as well as its execution efficiency, I PATH includes a fast bootstrapping algorithm to reconstruct a known set of paths. I PATH achieves a much higher reconstruction ratio in networks with relatively low packet delivery ratio and high routing dynamics.

The contributions of this work are the following.

• We observe high path similarity in a real-world sensor network. Based on this observation, we propose an iterative boosting algorithm for efficient path inference. We propose a lightweight hash function for efficient verification within I PATH. We further propose a fast bootstrapping algorithm to improve the inference capability as well as its execution efficiency.

• We propose an analytical model to calculate the successful reconstruction probability in various network conditions such as network scale, routing dynamics, packet losses, and node density.

• We implement I PATH and evaluate its performance using traces from large-scale WSN deployments as well as extensive simulations. I PATH achieves higher reconstruction ratio under different network settings compared to states of the art

II. LITERACTURE SERVEY

1) Monitoring heritage buildings with wireless sensor networks: The Torre Aquila deployment

Author: M. Ceriotti et al.,

Wireless sensor networks are un tethered infrastructures that are easy to deploy and have limited visual impact - a key asset in monitoring heritage buildings of artistic interest. This paper describes one such system deployed in Torre Aquila, a medieval tower in Trento (Italy). Our contributions range from the hardware to the graphical front-end. Customized hardware deals efficiently with high-volume vibration data, and specially-designed sensors acquire the building's deformation. Dedicated software services provide: i) data collection, to efficiently reconcile the diverse data rates and reliability needs of heterogeneous sensors; ii) data dissemination, to spread configuration changes and enable remote tasking; iii) time synchronization, with low memory demands. Unlike most deployments, built directly on the operating system, our entire software layer sits atop our Teeny LIME middleware. Based on 4 months of operation, we show that our system is an effective tool for assessing the tower's stability, as it delivers data reliably (with loss ratios <0.01%) and has an estimated lifetime beyond one year. [1]

2) Canopy closure estimates with Green Orbs: Sustainable sensing in the forest

Author: L. Mo et al.,

Motivated by the needs of precise forest inventory and real-time surveillance for ecosystem management, in this paper we present Green Orbs [2], a wireless sensor network system and its application for canopy closure estimates. Both the hardware and software designs of Green Orbs are tailored for sensing in wild environments without human supervision, including a firm weatherproof enclosure of sensor motes and a light-weight mechanism for node state monitoring and data collection. By incorporating a pre- deployment training process as well as a distributed calibration method, the estimates of canopy closure stay accurate and consistent against uncertain sensory data and dynamic environments. We have implemented a prototype system of Green Orbs and carried out multiple rounds of deployments. The evaluation results demonstrate that Green Orbs outperforms the conventional approaches for canopy closure estimates. Some early experiences are reported in this paper. [2]

3) A high throughput path metric for multi-hop wireless routing,

Author: D. S. J. D. Couto, D. Aguayo, J. Bicket, and R. Morris,

This paper presents the expected transmission count metric (ETX), which finds high-throughput paths on multi- hop wireless networks. ETX minimizes the expected total number of packet transmissions (including retransmissions) required to successfully deliver a packet to the ultimate destination. The ETX

metric incorporates the effects of link loss ratios, asymmetry in the loss ratios between the two directions of each link, and interference among the successive links of a path. In contrast, the minimum hop-count metric chooses arbitrarily among the different paths of the same minimum length, regardless of the often large differences in throughput among those paths, and ignoring the possibility that a longer path might offer higher throughput. This paper describes the design and implementation of ETX as a metric for the DSDV and DSR routing protocols, as well as modifications to DSDV and DSR which allow them to use ETX. Measurements taken from a 29-node 802.11b test-bed demonstrate the poor performance of minimum hop-count, illustrate the causes of that poor performance, and confirm that ETX improves performance. For long paths the throughput improvement is often a factor of two or more, suggesting that ETX will become more useful as networks grow larger and paths become longer. [5]

4) Path Zip: Packet path tracing in wireless sensornetworks,

Author: X. Lu, D. Dong, Y. Liu, X. Liao, and L. Shanshan,

In order to provide reliable data delivery and system management for large-scale wireless sensor networks (WSNs), tracing the route paths of packets in a lightweight manner is crucial and critical. Real-time path tracing technology enables us to observe every data transmission and analyze network dynamics in a fine-grained fashion. Due to resource constraints of WSNs, however, it is difficult, if not impossible, to integrate into each packet with its full path information. We attempt to capture such information with inserting a small and constant overhead into each packet. In this design, Path Zip, each sensor node performs lightweight hash-based computations to passively label every packet forwarded. Meanwhile, the sink extracts the label information so as to leverage the pre-knowledge on the network to compute the full packet path. Both topology-aware and geometry-assistant techniques are utilized by Path Zip in order to exploit different network knowledge and reduce the computation and storage overhead greatly. We conduct theoretical analysis and extensive simulations to evaluate the performance of our design. The results show that our method is effective to trace the full route path in large-scale WSNs, and outperforms the state-of-the-art methods. [7]

EXISTING SYSTEM:

- With the routing path of each packet, many measurement and diagnostic approaches are able to conduct effective management and protocol optimizations for deployed WSNs consisting of a large number of unattended sensor nodes. For example, PAD depends on the routing path information to build a Bayesian network for inferring the root causes of abnormal phenomena.
- Path information is also important for a network manager to effectively manage a sensor network. For example, given the per-packet path information, a network manager can easily find out the nodes with a lot of packets forwarded by them, i.e., network hop spots. Then, the manager can take actions to deal with that problem, such as deploying more nodes to that area and modifying the routing layer protocols.
- Furthermore, per-packet path information is essential to monitor the fine-grained per-link metrics. For example, most existing delay and loss measurement approaches assume that the routing topology is given as *a priori*.
- The time-varying routing topology can be effectively

obtained by per-packet routing path, significantly improving the values of existing WSN delay and loss tomography approaches.

DIS ADVANTAGES

From the above studies we conclude that these are the problems the existing system

- The growing network scale and the dynamic nature of wireless channel make WSNs become increasinglycomplex and hard to manage.
- The problem of existing approach is that its message overhead can be large for packets with long routing paths.
- Considering the limited communication resources of WSNs, this approach is usually not desirable in practice.

III. I PATH FOR WIRELESS SENSOR SYSTEM

- In this paper, we propose I PATH, a novel path inference approach to reconstruct routing paths at the sink side. Based on a real-world complex urban sensing network with all node generating local packets, we find a key observation: It is highly probable that a packet from node and *one of* the packets from ‘s parent will follow the same path starting from ‘s parent toward the sink. We refer to this observation as *high path similarity*.
- The basic idea of I PATH is to exploit high path similarity to iteratively infer long paths from short ones. I PATH starts with a known set of paths (e.g., the one-hop paths are already known) and performs path inference iteratively. During each iteration, it tries to infer paths one hop longer until no paths can be inferred.
- In order to ensure correct inference, I PATH needs to verify whether a short path can be used for inferring a long path. For this purpose, I PATH includes a novel design of a lightweight hash function. Each data packet attaches a hash value that is updated hop by hop. This *recorded hash value* is compared against the *calculated hash value* of an inferred path. If these two values match, the path is correctly inferred with a very high probability.
- In order to further improve the inference capability as well as its execution efficiency, I PATH includes a fast bootstrapping algorithm to reconstruct a known set ofpaths.

SYSTEM ARCHITECTURE:

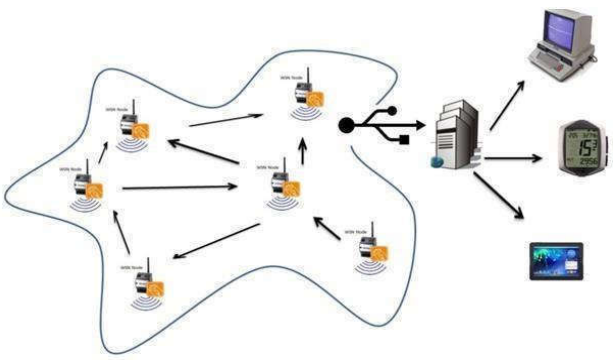


FIG 1-SYSTEM ARCHITECTURE

ADVANTAGES

- We observe high path similarity in a real-world sensor network.
- It’s an iterative boosting algorithm for efficient path inference.
- It’s a lightweight hash function for efficient verification within I PATH.
- The proposed system further proposes a fast bootstrapping algorithm to improve the inference capability as well as its execution efficiency.

I PATH achieves higher reconstruction ratio under different network settings

IV.IMPLEMENTATION

SOURCE

In this module, service provider browses the file; enter the file name and sends to the I PATH router. Service provider encrypts the data and send to the router.

I PATH ROUTER

In this module, router receives the file packets from the source, if packets size is greater than node BW then congestion occurs and then path inference will take place in order to find an alternative path. It takes another node and reaches the destination and load balancing takes place. When congestion occurs node band width can be increased.

RECEIVER

In this module, receiver receives the file. Calculates the time delay to reach the file from source to destination. Receiver stores the data details.

Data Flow diagram

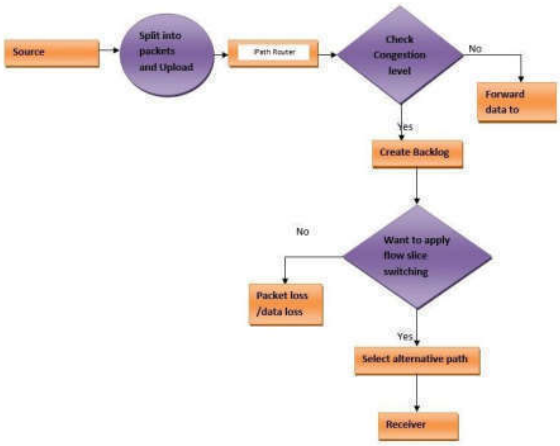


FIG-2 –DATA FLOW DIAGRAM

THE ITERATIVE BOOSTING ALGORITHM

Input: An initial set of packets P_{init} whose paths have been reconstructed and a set of other packets P_x
Output: The routing paths of packets

```
1: Procedure ITERATIVE-BOOSTING
2:     P_n ← P_init
3:     while P_n ≠ {} do
4:         P_nn ← {}
5:         for each packet k in P_n do
6:             for each packet i in P_x do
7:                 res = RECOVER(k, i)
8:                 if res ≡ True then
9:                     P_nn ← P_nn ∪ i
10:                    P_x ← P_x − i
11:        P_n ← P_nn
12: procedure RECOVER (k, i)
13:    if len(i) − len(k) ∉ {1,2} then
14:        return False
15:    if len(i) − len(k) ≡ 2
16:        if hash(o(i), p(i), path(k)) ≡ h(i) then
17:            path(i) ← (o(i), p(i), path(k)) // Case 2
18:            return True
19:        return False
20:    if len(i) − len(k) ≡ 1 then
21:        if hash(o(i), path(k)) ≡ h(i)
22:            path(i) ← (o(i), path(k))//Case 1
23:            return True
24:        if hash(o(i), p(i), path(k) − o(k)) ≡ h(i) then
25:            path(i) ← (o(i), p(i), path(k) − o(k)) // Case 3
26:            return True
27:        return False
```

V.CONCLUSION

In this paper, we propose I PATH, a novel path inference approach to reconstructing the routing path for each received packet. I PATH exploits the path similarity and uses the iterative boosting algorithm to reconstruct the routing path effectively. Furthermore, the fast bootstrapping algorithm provides an initial set of paths for the iterative algorithm. We formally analyze the reconstruction performance of I PATH as well as two related approaches. The analysis results show that I PATH achieves higher reconstruction ratio when the network setting varies. We also implement I PATH and evaluate its performance by a trace-driven study and extensive simulations. Compared to states of the art, I PATH achieves much higher reconstruction ratio under different network settings.

REFERENCES

[1] M. Ceriotti *et al.*, "Monitoring heritage buildings with wireless sensor networks: The Torre Aquila deployment," in *Proc. IPSN*, 2009, pp. 277–288.
[2] L. Mo *et al.*, "Canopy closure estimates with GreenOrbs: Sustainable sensing in the forest," in *Proc. SenSys*, 2009, pp. 99–112.
[3] X. Mao *et al.*, "CitySee: Urban CO2 monitoring with sensors," in *Proc. IEEE INFOCOM*, 2012, pp. 1611–1619.
[4] O. Gnawali, R. Fonseca, K. Jamieson, D. Moss, and P. Levis, "Collection tree protocol," in *Proc. SenSys*, 2009, pp. 1–14.
[5] D. S. J. D. Couto, D. Aguayo, J. Bicket, and R. Morris, "A highthroughput path metric for multi-hop wireless routing," in *Proc. MobiCom*, 2003, pp. 134–146.
[6] Z. Li, M. Li, J. Wang, and Z. Cao, "Ubiquitous data collection for mobile users in wireless sensor networks," in *Proc. IEEE INFOCOM*, 2011, pp. 2246–2254.
[7] X. Lu, D. Dong, Y. Liu, X. Liao, and L. Shanshan, "PathZip: Packet path tracing in wireless sensor networks," in *Proc. IEEE MASS*, 2012, pp. 380–388.
[8] M. Keller, J. Beutel, and L. Thiele, "How was your journey? Uncovering routing dynamics in deployed sensor networks with multi-hop network tomography," in *Proc. SenSys*, 2012, pp. 15–28.
[9] Y. Yang, Y. Xu, X. Li, and C. Chen, "A loss inference algorithm for wireless sensor networks to improve data reliability of digital ecosystems.," *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2126–2137, Jun. 2011.
[10] Y. Liu, K. Liu, and M. Li, "Passive diagnosis for wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 18, no. 4, pp. 1132–1144, Aug. 2010.
[11] W. Dong, Y. Liu, Y. He, T. Zhu, and C. Chen, "Measurement and analysis on the packet delivery performance in a large-scale sensor network," *IEEE/ACM Trans. Netw.*, 2013, to be published.
[12] J. Wang, W. Dong, Z. Cao, and Y. Liu, "On the delay performance analysis in a large-scale wireless sensor network," in *Proc. IEEE RTSS*, 2012, pp. 305–314.
[13] Y. Liang and R. Liu, "Routing topology inference for wireless sensor networks," *Comput. Commun. Rev.*, vol. 43, no. 2, pp. 21–28, 2013.
[14] Y. Gao *et al.*, "Domo: Passive per-packet delay tomography in wireless ad-hoc networks," in *Proc. IEEE ICDCS*, 2014, pp. 419–428.
[15] M. Lee, S. Goldberg, R. R. Kompella, and G. Varghese, "Fine-grained latency and loss measurements in the presence of reordering," in *Proc. ACM SIGMETRICS*, 2011, pp. 329– 340.
[16] Y. Shavitt and U. Weinsberg, "Quantifying the importance of vantage points distribution in internet topology measurements," in *Proc. IEEE INFOCOM*, 2009, pp. 792–800.
[17] M. Latapy, C. Magnien, and F. Oudraogo, "A radar for the internet," in *Proc. IEEE ICDMW*, 2008, pp. 901–908.
[18] I. Cunha, R. Teixeira, D. Veitch, and C. Diot, "Predicting and tracking internet path changes," in *Proc. SIGCOMM*, 2011, pp. 122–133.
[19] A. D. Jaggard, S. Kopparty, V. Ramachandran, and R. N. Wright, "The design space of probing algorithms for

network-performance measurement," in *Proc. SIGMETRICS*, 2013, pp. 105–116.
[20] L. Ma, T. He, K. K. Leung, A. Swami, and D. Towsley, "Identifiability of link metrics based on end-to-end path measurements," in *Proc. IMC*, 2013, pp. 391–404.
[21] Y. Gao *et al.*, "Pathfinder: Robust path reconstruction in large scale sensor networks with lossy links," in *Proc. IEEE ICNP*, 2013, pp. 1–10.
[22] A. Woo, T. Tong, and D. Culler, "Taming the underlying challenges of reliable multihop routing in sensor networks," in *Proc. SenSys*, 2003, pp. 14–27.
[23] Y. Gao *et al.*, "iPath: Path inference in wireless sensor networks," Tech. Rep., 2014 [Online]. Available: http://www.emnets.org/pub/ gaoyi/tech-ipath.pdf
[24] A. Liu and P. Ning, "TinyECC: A configurable library for elliptic curve cryptography in wireless sensor networks," in *Proc. IPSN*, 2008, pp. 245–256.
[25] V. Handziski, A. Köpke, A. Willig, and A. Wolisz, "TWIST: A scalable and reconfigurable testbed for wireless indoor experiments with sensor networks," in *Proc. REALMAN*, 2006, pp. 63–70.
[26] R. Lim, C. Walser, F. Ferrari, M. Zimmerling, and J. Beutel, "Distributed and synchronized measurements with FlockLab," in *Proc. SenSys*, 2012, pp. 373–374.
[27] Z. Li, M. Li, and Y. Liu, "Towards energy-fairness in asynchronous duty-cycling sensor networks," *Trans. Sensor Netw.*, vol. 10, no. 3, pp. 38:1–38:26, 2014.
[28] P. Levis, N. Lee, M. Welsh, and D. Culler, "TOSSIM: Accurate and scalable simulation of entire TinyOS applications," in *Proc. SenSys*, 2003, pp. 126–137.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Aggressive and Community Auditing System with Clear Agreement for Cloud Data

Naveena Vemula
PG Student
Department of CSE
Malla Reddy College of Engineering
Hyderabad, India
vemula.naveena1995@gmail.com

Subhashini Peneti
Assistant Professor
Department of CSE
Malla Reddy College of Engineering
Hyderabad, India
subhashinivalluru@gmail.com

***Abstract*— Cloud users no longer physically possess their data, so how to ensure the integrity of their outsourced data becomes a Challenging task. Recently proposed schemes such as "provable data possession" and "proofs of retrievability" are designed to address this problem, but they are designed to audit static archive data and therefore lack of data dynamics support. Moreover, threat models in these schemes usually assume an honest data owner and focus on detecting a dishonest cloud service provider despite the fact that clients may also misbehave. This paper proposes a public auditing scheme with data dynamics support and fairness arbitration of potential disputes. In particular, we design an index switcher to eliminate the limitation of index usage in tag computation in current schemes and achieve efficient handling of data dynamics.**

Keywords*— *Integrity, auditing, verifiability, arbitration, fairness.

I. INTRODUCTION

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the common use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation. Cloud computing consists of hardware and software resources made available on the Internet as managed third-party services. These services typically provide access to advanced software applications and high-end networks of server computers.

The goal of cloud computing is to apply traditional supercomputing, or high-performance computing power, normally used by military and research facilities, to perform tens of trillions of computations per second, in consumer-oriented applications such as financial portfolios, to deliver personalized information, to provide data storage or to power large, immersive computer games. The cloud computing uses networks of large groups of servers typically running low-cost consumer PC technology with specialized connections to spread data-processing chores across them. This shared IT infrastructure contains large pools of systems that are linked together. Often, virtualization techniques are used to maximize the power of cloud computing.

A. Characteristics and Services Models:

The salient characteristics of cloud computing based on the definitions provided by the National Institute of Standards and Terminology (NIST) are outlined below:

On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider[9].

Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).

B. Characteristics and Services Models:

The salient characteristics of cloud computing based on the definitions provided by the National Institute of Standards and Terminology (NIST) are outlined below:

On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.

Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).

Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location-independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or data center). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.

Rapid elasticity: Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be managed, controlled, and reported providing transparency for both the provider and consumer of the utilized service.

C. Services Models:

Cloud Computing comprises three different service models, namely Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). The three service models or layer are completed by an end user layer that encapsulates the end user perspective on cloud services. The model is shown in figure below. If a cloud user accesses services on the infrastructure layer, for instance, she can run her own applications on the resources of a cloud infrastructure and remain responsible for the support, maintenance, and security of these applications

herself. If she accesses a service on the application layer, these tasks are normally taken care of by the cloud service provider.

D. Benefits of cloud computing:

1. Achieve economies of scale – increase volume output or productivity with fewer people. Your cost per unit, project or product plummets.
2. Reduce spending on technology infrastructure. Maintain easy access to your information with minimal upfront spending. Pay as you go (weekly, quarterly or yearly), based on demand.
3. Globalize your workforce on the cheap**.** People worldwide can access the cloud, provided they have an Internet connection.
4. Streamline processes. Get more work done in less time with less people.
5. Reduce capital costs**.** There's no need to spend big money on hardware, software or licensing fees.
6. Improve accessibility. You have access anytime, anywhere, making your life so much easier!
7. Monitor projects more effectively. Stay within budget and ahead of completion cycle times.
8. Less personnel training is needed. It takes fewer people to do more work on a cloud, with a minimal learning curve on hardware and software issues.
9. Minimize licensing new software. Stretch and grow without the need to buy expensive software licenses or programs.
10. Improve flexibility. You can change direction without serious "people" or "financial" issues at stake.

II. LITERATURE SURVEY

This section provides different types of data Integrity auditing schemes:

A. Data Integrity Auditing Schemes:

1. Remote integrity checking
2. Demonstrating data possessing and uncheatable data transfer
3. Pors: Proofs of retrievability for large file
4. Provable data possession at untrusted stores
5. Compact proofs of retrievability

a) Remote integrity checking

This paper analyzes the problem of checking the integrity of files stored on remote servers. Since servers are prone to successful attacks by malicious hackers, the result of simple integrity checks run on the servers cannot be trusted. Conversely, downloading the files from the server to the verifying host is impractical. Two solutions are proposed, based on challenge-response protocols [1].

b) Demonstrating data possessing and uncheatable data transfer

We observe that a certain RSA-based secure hash function is homomorphic. We describe a protocol based on this hash function which prevents 'cheating' in a data transfer transaction, while placing little burden on the trusted third party that oversees the protocol. We also describe a cryptographic protocol based on similar principles, through which a prover can demonstrate possession of an arbitrary set of data known to the verifier. The verifier isn't required to have this data at hand during the protocol execution, but rather only a small hash of it. The protocol is also provably as secure as integer factoring[2].

c) Pors: Proofs of retrievability for large file

In this paper, we define and explore proofs of retrievability (PORs). A POR scheme enables an archive or back-up service (prover) to produce a concise proof that a user (verifier) can retrieve a target file *F*, that is, that the archive retains and reliably transmits file data sufficient for the user to recover *F* in its entirety. A POR may be viewed as a kind of cryptographic proof of knowledge (POK), but one specially designed to handle a *large* file (or bit string) *F*. We explore POR protocols here in which the communication costs, number of memory accesses for the prover, and storage requirements of the user (verifier) are small parameters essentially independent of the length of *F*. In addition to proposing new, practical POR constructions, we explore implementation considerations and optimizations that bear on previously explored, related schemes. In a POR, unlike a POK, neither the prover nor the verifier need actually have knowledge of *F*. PORs give rise to a new and unusual security definition whose formulation is another contribution of our work. We view PORs as an important tool for semi-trusted online archives. Existing cryptographic techniques help users ensure the privacy and integrity of files they retrieve. It is also natural, however, for users to want to verify that archives do not delete or modify files prior to retrieval. The goal of a POR is to accomplish these checks *without users having to download the files themselves*. A POR can also provide quality-of-service guarantees, i.e., show that a file is retrievable within a certain time bound[3][6].

d) Provable data possession at untrusted stores

We introduce a model for *provable data possession* (PDP) that allows a client that has stored data at an untrusted server to verify that the server possesses the original data without retrieving it. The model generates probabilistic proofs of possession by sampling random sets of blocks from the server, which drastically reduces I/O costs. The client maintains a constant amount of metadata to verify the proof. The challenge/response protocol transmits a small, constant amount of data, which minimizes network communication. Thus, the PDP model for remote data checking supports large data sets in widely-distributed storage system.

We present two provably-secure PDP schemes that are more efficient than previous solutions, even when compared with schemes that achieve weaker guarantees. In particular, the overhead at the server is low (or even constant), as opposed to linear in the size of the data. Experiments using our implementation verify the practicality of PDP and reveal that the

performance of PDP is bounded by disk I/O and not by cryptographic computation[4].

e) Compact proofs of retrievability

In a proof-of-retrievability system, a data storage center convinces a verifier that he is actually storing all of a client's data. The central challenge is to build systems that are both efficient and *provably* secure—that is, it should be possible to extract the client's data from any prover that passes a verification check. In this paper, we give the first proof-of-retrievability schemes with full proofs of security against *arbitrary* adversaries in the strongest model, that of Juels and Kaliski. Our first scheme, built from BLS signatures and secure in the random oracle model, has the *shortest query and response* of any proof-of-retrievability with public verifiability. Our second scheme, which builds elegantly on pseudorandom functions (PRFs) and is secure in the standard model, has the *shortest response* of any proof-of-retrievability scheme with private verifiability (but a longer query). Both schemes rely on homomorphic properties to aggregate a proof into one small authenticator value[5].

B. *Disadvantages of existing data integrity schemes*

a) Providing data dynamics support is the most challenging. This is because most existing auditing schemes intend to embed a block's index into its tag computation, which serves to authenticate challenged blocks. However, if we insert or delete a block, block indices of all subsequent blocks will change, then tags of these blocks have to be re-computed. This is unacceptable because of its high computation overhead.

b) Current research usually assumes an honest data owner in their security models, which has an inborn inclination toward cloud users. However, the fact is, not only the cloud, but also cloud users, have the motive to engage in deceitful behaviors.

c) In Existing System no integrity auditing scheme with public verifiability, efficient data dynamics and fair disputes arbitration.

d) Existing system has the limitation of index usage in tag computation

e) In Existing System tag re-computation caused by block update operations

f) In Existing System both clients and the CSP potentially may misbehave during auditing and data update.

III. PROPOSED WORK:DYNAMIC PUBLIC AUDITING SCHEME

The public auditing system is designed based on tag index (used for tag computation) and block index (indicate block position), and rely an index switcher to keep a mapping between them. Upon each update operation, we allocate a new tag index for the operating block and update the mapping between tag indices and block indices. Such a layer of indirection between block indices and tag indices enforces block authentication and avoid stag re-computation of blocks after the operation position simultaneously[7][8].

A. *The public auditing scheme consist of different participants*:

1. User
2. Third Party Auditor(TPA)
3. Third Party Arbitrator(TPAR)
4. Cloud Service Provider(CSP)

a) user

The data owner performs the following actions:

User performs User registration and user login details.

- Every User need to register while accessing to the cloud.
- Every user will be activated by the Cloud.
- After Cloud activated, every user need to provide public key to login the user home.
- Public key will be provided to the user by third party auditor.
- User can view file details and can insert, modify and delete the file with help of TPAR.
- User will have the TPAR message whenever the user update the file

b) Third Party Auditor(TPA)

It acts as semi-cloud.

- TPA Provide public key for every user to access the user home page.
- After cloud given auditing proof then only TPA can audit all files

c) Third Party Arbitrator(TPAR)

- It acts as fair dispute for users and cloud.
- Intimate the files message, each time user insert, modify, delete files to cloud.
- Send TPAR message to user and cloud.

d) Cloud Service Provider(CSP)

- Activate data for user.
- Cloud sends storage auditing proof for all files to TPA.
- Cloud can view the user downloaded files from cloud.
- Cloud will have the TPAR message whenever the user updates the file.

As a result, the efficiency of handling data dynamics is greatly enhanced. Furthermore and important, in a public auditing scenario, a data owner always delegates his auditing tasks to a TPA who is trusted by the owner but not necessarily by the cloud. Our work also adopts the idea of signature exchange to ensure the metadata correctness and protocol fairness, and we concentrate on combining efficient data dynamics support and fair dispute arbitration into a single auditing scheme.

To address the fairness problem in auditing, we introduce a third-party arbitrator (TPAR) into our threat model, which is a professional institute for conflicts arbitration and is trusted and payed by both data owners and the CSP. Since a TPA can be viewed as a delegator of the data owner and is not necessarily trusted by the CSP, we differentiate between the roles of auditor

and arbitrator. Moreover, we adopt the idea of signature exchange to ensure metadata correctness and provide dispute arbitration, where any conflict about auditing or data update can be fairly arbitrated. Generally, this paper proposes a new auditing scheme to address the problems of data dynamics support, public verifiability and dispute arbitration simultaneously.

B. *Public Auditing Scheme Architecture*

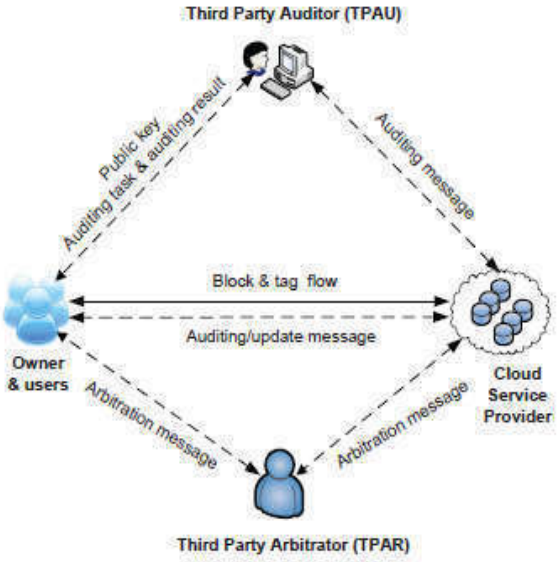


Fig. 1. The system and threat model.

As illustrated in Fig. 1, the system model involves four different entities: the data owner/cloud user, who has a large amount of data to be stored in the cloud, and will dynamically update his data (e.g., insert, delete or modify a data block) in the future; the cloud service provider (CSP), who has massive storage space and computing power that users do not possess, stores and manages user's data and related metadata (i.e., the tag set and the index switcher); the third party auditor (TPAU) is similar to the role of TPA in existing schemes, who is a public verifier with expertise and capabilities for auditing, and is trusted and payed by the data owner (but not necessarily trusted by the cloud) to assess the integrity of the owner's remotely stored data; the third party arbitrator (TPAR), who is a professional institute for conflict arbitration and trusted by both the owner and the CSP, which is different to the role of TPAU. Cloud users rely on the CSP for data storage and maintenance, and they may access and update their data. To alleviate their burden, cloud users can delegate auditing tasks to the TPAU, who periodically performs the auditing and honestly reports the result to users. Additionally, cloud users may perform auditing tasks themselves if necessary. For potential disputes between the auditor and the CSP, the TPAR can fairly settle the disputes on proof verification or data update. Note in following sections, we may use the terms "TPAU" and "auditor" interchangeably, so are the terms "TPAR" and "arbitrator".

C. *Advantages Of Proposed System*

- The proposed system solves the data dynamics problem in auditing by introducing an index switcher to keep a mapping between block indices and tag indices, and eliminate the passive effect of block indices in tag computation without incurring much overhead[11].
- The proposed system extend the threat model in current research to provide dispute arbitration, which is of great significance and practicality for cloud data auditing, since most existing schemes generally assume an honest data owner in their threat models[10].
- The proposed system provides fairness guarantee and dispute arbitration in our scheme, which ensures that both the data owner and the cloud cannot misbehave in the auditing process or else it is easy for a third-party arbitrator to find out the cheating party[11][12].

IV. PERFORMANCE EVALUATION

Our method public auditing scheme is implemented using C language on a Linux system. We measure the performance of our auditing scheme for different block sizes.

Table 1. Peformance of public auditing scheme

Block size	Tag cost(s)
128	0
64	5
32	10
16	15
8	20
4	25
2	30

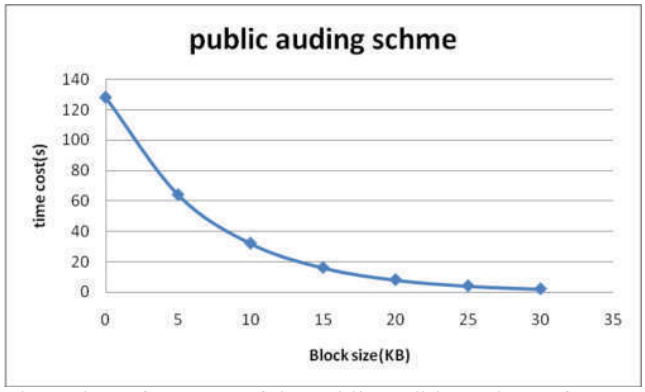


Fig2. The peformance of the public auditing scheme for different block sizes

V. CONCLUSION

The aim of this paper is to provide an integrity auditing scheme with public verifiability, efficient data dynamic sand fair disputes arbitration. To eliminate the limitation of index usage in tag computation and efficiently support data dynamics, we differentiate between block indices and tag indices, and devise an index switcher to keep block-tag index mapping to avoid tag re-computation caused by block update operations, which incurs limited additional overhead, as shown in our performance evaluation. Meanwhile, since both clients and the CSP potentially may misbehave during auditing and data update, we extend the existing threat model in current research to provide fair arbitration for solving disputes between clients and the CSP, which is of vital significance for the deployment and promotion of auditing schemes in the cloud environment. We achieve this by designing arbitration protocols based on the

idea of exchanging metadata signatures upon each update operation. Our experiments demonstrate the efficiency of our proposed scheme, whose overhead for dynamic update and dispute arbitration are reasonable.

REFERENCES

[1] Y. Deswarte, J.-J. Quisquater, and A. Sa¨ıdane, "Remote integrity checking," in *Proc. 5th Working Conf. Integrity and Intl Control inInformation Systems*, 2004, pp. 1–11.

[2] D. L. Gazzoni Filho and P. S. L. M. Barreto, "Demonstrating data possession and uncheatable data transfer." *IACR Cryptology ePrint Archive, Report 2006/150*, 2006.

[3] A. Juels and B. S. Kaliski Jr, "Pors: Proofs of retrievability for large files," in *Proc. 14th ACM Conf. Computer and Comm. Security (CCS07)*, 2007, pp. 584–597.

[4] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable data possession at untrusted stores," in *Proc. 14th ACM Conf. Computer and Comm. Security (CCS07)*, 2007, pp. 598–609.

[5] H. Shacham and B. Waters, "Compact proofs ofretrievability," in *Proc. 14th Intl Conf. Theory and Application of Cryptology and Information Security: Advances in Cryptology (ASIACRYPT 08)*, 2008, pp. 90–107.

[6] Q. Wang, C. Wang, J. Li, K. Ren, and W. Lou, "Enabling public verifiability and data dynamics for storage security in cloud computing," in *Proc. 14th European Conf. Research in Computer Security (ESORICS 08)*, 2009, pp. 355–370.

[7] M. A. Shah, R. Swaminathan, and M. Baker, "Privacy preserving audit and extraction of digital contents." *IACRCryptology ePrint Archive, Report 2008/186*, 2008.

[8] C. Wang, K. Ren, W. Lou, and J. Li, "Toward publicly auditable secure cloud data storage services," *Network, IEEE*, vol. 24, no. 4, pp. 19–24, 2010.

[9] C. Erway, A. K¨upc¸ ¨u, C. Papamanthou, and R. Tamassia, "Dynamic provable data possession," in *Proc. 16th ACM Conf. Computer and Comm. Security (CCS 09)*, 2009, pp. 213–222.

[10] Y. Zhu, H.Wang, Z. Hu, G.-J. Ahn, H. Hu, and S. S. Yau, "Dynamic audit services for integrity verification of outsourced storages in clouds," in *Proc. ACM Symp. Applied Computing (SAC 11)*, 2011, pp. 1550–1557.

[11] Q. Zheng and S. Xu, "Fair and dynamic proofs of retrievability," in *Proc. 1st ACM Conf. Data and Application Security and Privacy (CODASPY 11)*, 2011, pp. 237–248.

[12] A. K¨upc¸ ¨u, "Official arbitration with secure cloud storage application,"

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

IOT in Rural India

Anil Kumar Vutikuri,
PG Student,
Dept. of CSE,
anil.suryapet@gmail.com
Malla Reddy College of Engineering

Shruthi L,
PG Student,
Dept. of CSE,
shruthi.vsp19@gmail.com
Malla Reddy College of Engineering

Ch.Vijaya Kumari,
Associate Professor,
Dept. of CSE,
nagvijji@gmail.com
Malla Reddy College of Engineering

Abstract: Rural India is the base of our Country. In Rural India, Harvesting and Agriculture is top bread-wining activity. In India about 70% of population depends upon farming and one third of the nation's capital comes from farming. Issues concerning agriculture have been always hindering the development of the country. The only solution to this problem is smart agriculture by modernizing the current traditional methods of agriculture. Hence the project aims at making agriculture smart using automation and IoT technologies. The highlighting features of this project includes smart GPS based remote controlled robot to perform tasks like weeding, spraying, moisture sensing, bird and animal scaring, keeping vigilance, etc. Secondly it includes smart irrigation with smart control and intelligent decision making based on accurate real time field data. Thirdly, smart warehouse management which includes temperature maintenance, humidity maintenance and theft detection in the warehouse. Controlling of all these operations will be through any remote smart device or computer connected to Internet and the operations will be performed by interfacing sensors, Wi-Fi or ZigBee modules, camera and actuators with micro-controller and raspberry pi.

Keywords: IoT, automation, Wi-Fi

I. INTRODUCTION

Agriculture is considered as the basis of life for the human species as it is the main source of food grains and other raw materials. It plays vital role in the growth of country's economy. It also provides large ample employment opportunities to the people. Growth in agricultural sector is necessary for the development of economic condition of the country. Unfortunately, many farmers still use the traditional methods of farming which results in low yielding of crops and fruits. But wherever automation had been implemented and human beings had been replaced by automatic machineries, the yield has been improved. Hence there is need to implement modern science and technology in the agriculture sector for increasing the yield. Most of the papers signifies the use of wireless sensor network which collects the data from different types of sensors and then send it to main server using wireless protocol. The collected data provides the information about different environmental factors which in turns helps to monitor the system. Monitoring environmental factors is not enough and complete solution to improve the yield of the crops. There are number of other factors that affect the productivity to great extent. These factors include attack of insects and pests which can be controlled by spraying the crop with proper insecticide and pesticides. Secondly, attack of wild animals and birds when the crop grows up. There is also possibility of thefts when crop is at the stage of harvesting. Even after harvesting, farmers also face problems in storage of harvested crop. So, in order to provide solutions to all such problems, it is necessary to develop integrated system which will take care of all factors affecting the productivity in every stages like; cultivation, harvesting and post harvesting storage. This paper therefore proposes a system which is useful in monitoring the field data as well as controlling the field operations which provides the flexibility. The paper aims at making agriculture smart using automation and IoT technologies. The highlighting features of this paper includes smart GPS based remote controlled robot to perform tasks like; weeding, spraying, moisture sensing, bird and animal scaring, keeping vigilance, etc. Secondly, it includes smart irrigation with smart control based on real time field data. Thirdly, smart warehouse management which includes; temperature maintenance, humidity maintenance and theft detection in the warehouse. Controlling of all these operations will be through any remote smart device or computer connected to Internet and the operations will be performed by interfacing sensors, Wi-Fi or ZigBee modules, camera and actuators with micro-controller and raspberry pi.

II. LITERATURE REVIEW

The newer scenario of decreasing water tables, drying up of rivers and tanks, unpredictable environment present an urgent need of proper utilization of water. To cope up with this use of temperature and moisture sensor at suitable locations for monitoring of crops is implemented in. [1] An algorithm developed with threshold values of temperature and soil moisture can be programmed into a microcontroller-based gateway to control water quantity. The system can be powered by photovoltaic panels and can have a duplex communication link based on a cellular-Internet interface that allows data inspection and irrigation scheduling to be programmed through a web page. [2] The technological development in Wireless Sensor Networks made it possible to use in monitoring and control of greenhouse parameter in precision agriculture. [3] After the research in the agricultural field, researchers found that the yield of agriculture is decreasing day by day. However, use of technology in the field of agriculture

plays important role in increasing the production as well as in reducing the extra man power efforts. Some of the research attempts are done for betterment of farmers which provides the systems that use technologies helpful for increasing the agricultural yield.

A remote sensing and control irrigation system using distributed wireless sensor network aiming for variable rate irrigation, real time in field sensing, controlling of a site specific precision linear move irrigation system to maximize the productivity with minimal use of water was developed by Y. Kim . The system described details about the design and instrumentation of variable rate irrigation, wireless sensor network and real time in field sensing and control by using appropriate software. The whole system was developed using five in field sensor stations which collects the data and send it to the base station using global positioning system (GPS) where necessary action was taken for controlling irrigation according to the database available with the system. The system provides a promising low cost wireless solution as well as remote controlling for precision irrigation. [4]

In the studies related to wireless sensor network, researchers measured soil related parameters such as temperature and humidity. Sensors were placed below the soil which communicates with relay nodes by the use of effective communication protocol providing very low duty cycle and hence increasing the life time of soil monitoring system. The system was developed using microcontroller, universal asynchronous receiver transmitter (UART) interface and sensors while the transmission was done by hourly sampling and buffering the data, transmit it and then checking the status messages. The drawbacks of the system were its cost and deployment of sensor under the soil which causes attenuation of radio frequency (RF) signals. [5]

III. SYSTEM OVERVIEW

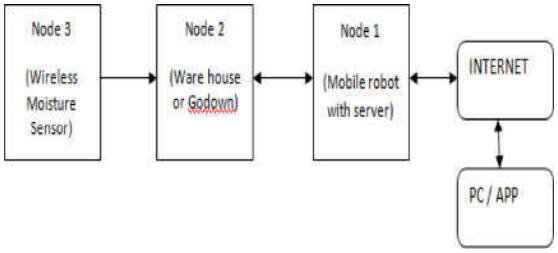


Figure 1: System overview

The paper consist of four sections; node1, node2, node3 and PC or mobile app to control system. In the present system, every node is integration with different sensors and devices and they are interconnected to one central server via wireless communication modules. The server sends and receives information from user end using internet connectivity. There are two modes of operation of the system; auto mode and manual mode. In auto mode system takes its own decisions and controls the installed devices whereas in manual mode user can control the operations of system using android app or PC commands.

IV. ARCHITECTURE OF THE SYSTEM

Node 1:

Node1 is GPS based mobile robot which can be controlled remotely using computer as well as it can be programmed so as to navigate autonomously within the boundary of field using the co-ordinates given by GPS module.

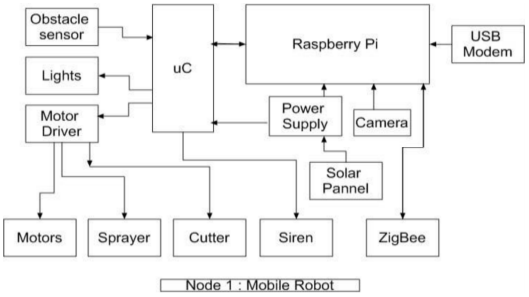


Figure 2: Node 1

The Remote controlled robot have various sensors and devices like camera, obstacle sensor, siren, cutter, sprayer and using them it will perform tasks like; Keeping vigilance, Bird and animal scaring, Weeding, and Spraying

Node 2:

Node2 will be the warehouse. It consists of motion detector, light sensor, humidity sensor, temperature sensor, room heater, cooling fan altogether interfaced with AVR microcontroller. Motion detector will detect the motion in the room when security mode will be ON and on detection of motion, it will send the alert signal to user via Raspberry pi and thus providing theft detection.

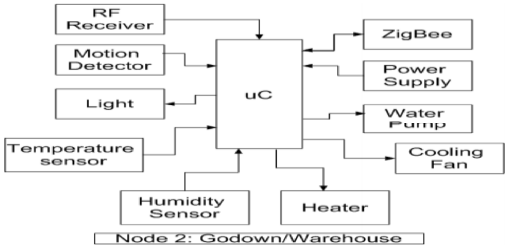


Figure 3: Node 2

Temperature sensor and Humidity sensor senses the temperature and humidity respectively and if the value crosses the threshold then room heater or cooling fan will be switched ON/OFF automatically providing temperature and humidity maintenance.Node2 will also controls water pump depending upon the soil moisture data sent by node3.

Node 3:

Node3 is a smart irrigation node with features like ; Smart control of water pump based on real time field data i.e. automatically turning on/off the pump after attaining the required soil moisture level in auto mode, Switching water pump on/off remotely via mobile or computer in manual mode, and continuous monitoring of soil moisture.

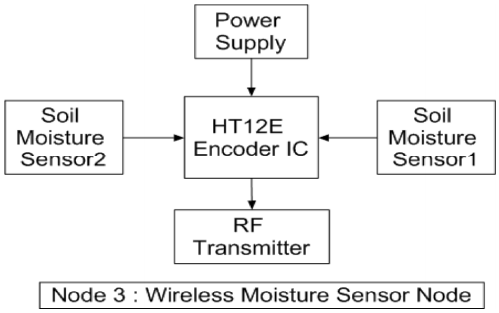


Figure 4: Node 3

In node3, moisture sensor transmits the data using HT12E Encoder IC and a RF transmitter. The transmitted data is received by node2 and there it is processed by microcontroller in order to control the operation of water pump.

Hardware used:

a) AVR Microcontroller Atmega 16/32:

The microcontroller used is, Low-power AVR® 8-bit Microcontroller, having 8K Bytes of In-System Self-programmable Flash program memory, Programmable Serial USART, 8-channel, 10-bit ADC, 23 Programmable I/O Lines.

b) ZigBee Module:

ZigBee is used for achieving wireless communication between Node1 and Node2. The range for Zigbee is roughly 50 meters and it can be increased using high power modules or by using network of modules. It operates on 2.4 GHz frequency. Its power consumption is very low and it is less expensive as compared to other wireless modules like Wi-Fi or Bluetooth. It is usually used to establish wireless local area networks.

c) Temperature Sensor LM35:

The LM35 is precision IC temperature sensor. Output voltage of LM35 is directly proportional to the Centigrade/Celsius of temperature. The LM35 does not need external calibration or trimming to provide accurate temperature range. It is very low cost sensor. It has low output impedance and linear output. The operating temperature range for LM35 is −55˚ to +150˚C. With rise in temperature, the output voltage of the sensor increases linearly and the value of voltage is given to the microcontroller which is multiplied by the conversion factor in order to give the value of actual temperature.

d) Moisture sensor:

Soil moisture sensor measures the water content in soil. It uses the property of the electrical resistance of the soil. The relationship among the measured property and soil moisture is calibrated and it may vary depending on environmental factors such as temperature, soil type, or electric conductivity. Here, It is used to sense the moisture in field and transfer it to microcontroller in order to take controlling action of switching water pump ON/OFF.

Humidity sensor:

The DHT11 is a basic, low-cost digital temperature and humidity sensor. It gives out digital value and hence there is no need to use conversion algorithm at ADC of the microcontroller and hence we can give its output directly to data pin instead of ADC. It has a capacitive sensor for measuring humidity. The only real shortcoming of this sensor is that one can only get new data from it only after every 2 seconds.

e) Obstacle sensor (Ultra-Sonic):

The ultra-sonic sensor operates on the principle of sound waves and their reflection property. It has two parts; ultra-sonic transmitter and ultra-sonic receiver. Transmitter transmits the 40 KHz sound wave and receiver receives the reflected 40 KHz wave and on its reception, it sends the electrical signal to the microcontroller. The speed of sound in air is already known.

Hence from time required to receive back the transmitted sound wave, the distance of obstacle is calculated. Here, it is used for obstacle detection in case of mobile robot and as a motion detector in ware house for preventing thefts. The ultra-sonic sensor enables the robot to detect and avoid obstacles and also to measure the distance from the obstacle. The range of operation of ultra-sonic sensor is 10 cm to 30 cm.

f) Raspberry Pi :

The Raspberry Pi is small pocket size computer used to do small computing and networking operations. It is the main element in the field of internet of things. It provides access to the internet and hence the connection of automation system with remote location controlling device becomes possible. Raspberry Pi is available in various versions. Here, model Pi 2 model B is used and it has quad-core ARM Cortex-A53 CPU of 900 MHz, and RAM of 1GB. it also has: 40 GPIO pins, Full HDMI port, 4 USB ports, Ethernet port, 3.5mm audio jack, video Camera interface (CSI), the Display interface (DSI), and Micro SD card slot.

Softwares used:

a) AVR Studio Version 4:

It is used to write, build, compile and debug the embedded c program codes which are needed to be burned in the microcontroller in order to perform desired operations. This software directly provides .hex file which can be easily burned into the microcontroller.

b) Proteus 8 Simulator:

Proteus 8 is one of the best simulation software for various circuit designs of microcontroller. It has almost all microcontrollers and electronic components readily available in it and hence it is widely used simulator.

It can be used to test programs and embedded designs for electronics before actual hardware testing. The simulation of programming of microcontroller can also be done in Proteus. Simulation avoids the risk of damaging hardware due to wrong design.

c) Dip Trace:
Dip race is EDA/CAD software for creating schematic diagrams and printed circuit boards. The developers provide multi-lingual interface and tutorials (currently available in English and 21 other languages). DipTrace has 4 modules: Schematic Capture Editor, PCB Layout Editor with built-in shape-based auto router and 3D Preview & Export, Component Editor, and Pattern Editor.

d) SinaProg:
SinaProg is a Hex downloader application with AVR Dude and Fuse Bit Calculator. This is used to download code/program and to set fuse bits of all AVR based microcontrollers.

e) Raspbian Operating System:
Raspbian operating system is the free and open source operating system which Debian based and optimized for Raspberry Pi. It provides the basic set of programs and utilities for operating Raspberry Pi. It comes with around 35,000 packages which are pre-compiled softwares that are bundled in a nice format for hustle free installation on Raspberry Pi.It has good community of developers which runs the discussion forms and provides solutions to many relevant problems. However, Raspbian OS is still under consistent development with an main focus on improving the performance and the stability of as many Debian packages as possible.

V. EXPERIMENTATION AND RESULTS



Figure 5: experimental setup for Node1

As shown in figure 5,experimental setup for node1 consists of mobile robot with central server, GPS module, camera and other sensors.All sensors are successfully interfaced with microcontroller and the microcontroller is interfaced with the raspberry pi.GPS and camera are also connected to raspberry pi. Test results shows that the robot can be controlled remotely using wireless transmission of PC commands to R-Pi.R-Pi forwards the commands to microcontroller and micocontroller gives signals to motor driver in order to drive the Robot.GPS module provides the co-ordinates for the location of the robot

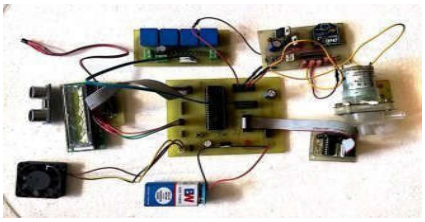


Figure 6: experimental setup for Node2

As shown in above figure, node2 consists of motion detector, temperature sensor, humidity senor, cooling fan, water pump, etc. connected to the microcontroller board.

The sensors gives input to the controller and according to that microcontroller controls the devices in auto mode and also sends the value of sensors to R-Pi and R-Pi forwards it to user's smart device using internet. Test results shows that when temperature level increases above preset threshold level then cooling fan is started automatically in auto mode.

The water pump also gets turned ON if moisture level goes below fixed threshold value. In manual mode, microcontroller receives the controlling signals from R-Pi through ZigBee and accordingly takes the controlaction.

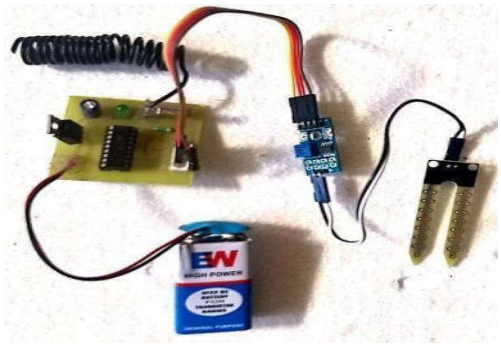


Figure 7: experimental setup for Node3

As shown in above figure, node3 consists of a moisture sensor connected to HT12E. Moisture sensor transmits the data using HT12E Encoder IC and a RF transmitter to the Node2 where it is processed by microcontroller and accordingly water pump is switched ON/OFF.

VI. CONCLUSION

The sensors and microcontrollers of all three Nodes are successfully interfaced with raspberry pi and wireless communication is achieved between various Nodes.

All observations and experimental tests proves that project is a complete solution to field activities, irrigation problems, and storage problems using remote controlled robot, smart irrigation system and a smart warehouse management system respectively. Implementation of such a system in the field can definitely help to improve the yield of the crops and overall production.

ACKNOWLEDGMENT

I am sincerely thankful to all my teachers for their guidance for my seminar. Without their help it was tough job for me to accomplish this task. I am especially very thankful to my guide **Dr. R.S. Kawitkar** for his consistent guidance, encouragement and motivation throughout the period of this work. I also want to thank our Head of the Department (E&TC) **Dr. M. B. Mali** for providing me all necessary facilities.

REFERENCES

[1] S. R. Nandurkar, V. R. Thool, R. C. Thool, "Design and Development of Precision Agriculture System Using Wireless Sensor Network", IEEE International Conference on Automation, Control, Energy and Systems (ACES), 2014

[2] JoaquínGutiérrez, Juan Francisco Villa-Medina, Alejandra Nieto-Garibay, and Miguel Ángel Porta-Gándara, "Automated Irrigation System Using a Wireless Sensor Network and GPRS Module",IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, 0018-9456,2013

[3] Dr. V .Vidya Devi,G. Meena Kumari, "Real- Time Automation and Monitoring System for Modernized Agriculture" ,International Journal of Review and Research in Applied Sciences and Engineering (IJRRASE) Vol3 No.1. PP 7-12, 2013

[4] Y. Kim, R. Evans and W. Iversen, "Remote Sensing and Control of an Irrigation System Using a Distributed Wireless Sensor Network", IEEE Transactions on Instrumentation and Measurement, pp. 1379–1387, 2008.

[5] Q. Wang, A. Terzis and A. Szalay, "A Novel Soil Measuring Wireless Sensor Network", IEEE Transactions on Instrumentation and Measurement, pp. 412–415, 2010

[6] Yoo, S.; Kim, J.; Kim, T.; Ahn, S.; Sung, J.; Kim, D. A2S: Automated agriculture system based on WSN. In ISCE 2007. IEEE International Symposium on Consumer Electronics, 2007, Irving,TX, USA, 2007

[7] Arampatzis, T.; Lygeros, J.; Manesis, S. A survey of applications of wireless sensors and Wireless Sensor Networks. In 2005 IEEE International Symposium on Intelligent Control & 13th Mediterranean Conference on Control and Automation. Limassol, Cyprus, 2005, 1-2, 719-724

[8] Orazio Mirabella and Michele Brischetto, 2011. "A Hybrid Wired/Wireless Networking Infrastructure for Greenhouse Management", IEEE transactions on instrumentation and measurement, vol. 60, no. 2, pp 398-407.

[9] N. Kotamaki and S. Thessler and J. Koskiaho and A. O. Hannukkala and H. Huitu and T. Huttula and J. Havento and M. Jarvenpaa(2009). "Wireless in-situ sensor network for agriculture and water monitoring on a river basin scale in Southern Finland: evaluation from a data users perspective". Sensors 4, 9: 2862-2883. doi:10.3390/s90402862 2009.

[10] Liu, H.; Meng, Z.; Cui, S. A wireless sensor network prototype for environmental monitoring in greenhouses. International Conference on Wireless Communications, Networking and Mobile Computing (WiCom 2007), Shangai, China; 21-25 September 2007.

[11] Baker, N. ZigBee and bluetooth - Strengths and weaknesses for industrial applications. Comput. Control. Eng. 2005, 16,20-25.

[12] IEEE, Wireless medium access control (MAC) and physical layer (PHY) specifications for lowrate wireless personal area networks (LR-WPANs). In The Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2003.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

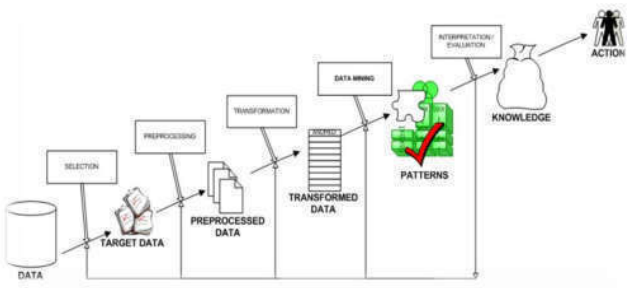
BOOSTER IN HIGH DIMENSIONAL DATA CLASSIFICATION

VISHWESHWAR KOKKULA
Dept of CSE
Malla Reddy College of Engineering
Maisammaguda, Dullapally
kokkulavishweshwar@gmail.com

Dr. S. Moses Dian
Dept of CSE
Malla Reddy College of Engineering
Maisammaguda, Dullapally
mosesdian_cse@mrce.in

Abstract— **Classification problems in high dimensional data with a small number of observations are becoming more common especially in microarray data. During the last two decades, lots of efficient classification models and feature selection (FS) algorithms have been proposed for higher prediction accuracies. However, the result of an FS algorithm based on the prediction accuracy will be unstable over the variations in the training set, especially in high dimensional data. This paper proposes a new evaluation measure Q-statistic that incorporates the stability of the selected feature subset in addition to the prediction accuracy. Then, we propose the Booster of an FS algorithm that boosts the value of the Q-statistic of the algorithm applied. Empirical studies based on synthetic data and 14 microarray data sets show that Booster boosts not only the value of the Q-statistic but also the prediction accuracy of the algorithm applied unless the data set is intrinsically difficult to predict with the given algorithm.**

INTRODUCTION



Structure of Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases..

Data mining consists of five major elements:

1) Extract, transform, and load transaction data onto the data warehouse system.
2) Store and manage the data in a multidimensional database system.
3) Provide data access to business analysts and information technology professionals.
4) Analyze the data by application software.
5) Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical

dataset (where k=1). Sometimes called the k-nearest neighbor technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

LITERATURE SURVEY

Biomarker discovery is an important topic in biomedical applications of computational biology, including applications such as gene and SNP selection from high-dimensional data. However, robustness of biomarkers is an important issue, as it may greatly influence subsequent biological validations. First contribution is a general framework for the analysis of the robustness of a biomarker selection algorithm. Secondly, they conducted a large-scale analysis of the recently introduced concept of ensemble feature selection, where multiple feature selections are combined in order to increase the robustness of the final set of selected features. We focus on selection methods that are embedded in the estimation of support vector machines (SVMs). SVMs are powerful classification models that have shown state-of-the-art performance on several diagnosis and prognosis tasks on biological data. Their feature selection extensions also offered good results for gene selection tasks. they show that the robustness of SVMs for biomarker discovery can be substantially increased by using ensemble feature selection techniques, while at the same time improving upon classification performances. The proposed methodology is evaluated on four microarray datasets showing increases of up to almost 30% in robustness of the selected biomarkers, along with an improvement of ~15% in classification performance. The stability improvement with ensemble methods is particularly noticeable for small signature sizes (a few tens of genes), which is most relevant for the design of a diagnosis or prognosis model from a gene signature [1].

Storing and using specific instances improves the performance of several supervised learning algorithms. These include algorithms that learn decision trees, classification rules, and distributed networks. However, no investigation has analyzed algorithms that use only specific instances to solve incremental learning tasks. In this paper, we describe a framework and methodology, called instance-based learning, that generates classification predictions using only specific instances. Instance-based learning algorithms do not maintain a set of abstractions derived from specific instances. This approach extends the nearest neighbor algorithm, which has large storage requirements. We describe how storage requirements can be significantly reduced with, at most, minor sacrifices in learning rate and classification accuracy. While the storage-reducing algorithm performs well on several real-world databases, its performance degrades rapidly with the level of attribute noise in training instances. Therefore, we extended it with a significance test to distinguish noisy instances. This extended algorithm's performance degrades gracefully with increasing noise levels and compares favorably with a noise-tolerant decision tree algorithm [2].

Diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, is clinically heterogeneous: 40% of patients respond well to current therapy and have prolonged survival, whereas the remainder succumb to the disease. [3] proposed that this variability in natural history reflects unrecognized molecular heterogeneity in the tumours. Using DNA microarrays, They have conducted a systematic characterization of gene expression in B-cell malignancies and show that there is diversity in gene expression among the tumours of DLBCL patients, apparently reflecting the variation in tumour proliferation rate, host response and differentiation state of the tumour. They identified two molecularly distinct forms of DLBCL which had gene expression patterns indicative of different stages of B-cell differentiation. One type expressed genes characteristic of germinal centre B cells ('germinal centre B-like DLBCL'); the second type expressed genes normally induced during in vitro activation of peripheral blood B cells ('activated B-like DLBCL'). Patients with germinal centre B-like DLBCL had a significantly better overall survival than those with activated B-like DLBCL. The molecular classification of tumours on the basis of gene

expression can thus identify previously undetected and clinically significant subtypes of cancer.

Oligonucleotide arrays can provide a broad picture of the state of the cell, by monitoring the expression level of thousands of genes at the same time. It is of interest to develop techniques for extracting useful information from the resulting data sets. [4] report the application of a two-way clustering method for analyzing a data set consisting of the expression patterns of different cell types. Gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. An efficient two-way clustering algorithm was applied to both the genes and the tissues, revealing broad coherent patterns that suggest a high degree of organization underlying gene expression in these tissues. Coregulated families of genes clustered together, as demonstrated for the ribosomal proteins. Clustering also separated cancerous from noncancerous tissue and cell lines from in vivo tissues on the basis of subtle distributed patterns of genes even when expression of individual genes varied only slightly between the tissues. Two-way clustering thus may be of use both in classifying genes into functional groups and in classifying tissues based on gene expression.

Early detection of ventricular fibrillation (VF) is crucial for the success of the defibrillation therapy in automatic devices. A high number of detectors have been proposed in [5] based on temporal, spectral, and time–frequency parameters extracted from the surface electrocardiogram (ECG), showing always a limited performance. The combination ECG parameters on different domain (time, frequency, and time–frequency) using machine learning algorithms has been used to improve detection efficiency. However, the potential utilization of a wide number of parameters benefiting machine learning schemes has raised the need of efficient feature selection (FS) procedures. In this study, we propose a novel FS algorithm based on support vector machines (SVM) classifiers and bootstrap resampling (BR) techniques. We define a backward FS procedure that relies on evaluating changes in SVM performance when removing features from the input space. This evaluation is achieved according to a nonparametric statistic based on BR. After simulation studies, we benchmark the performance of our FS algorithm in AHA and MIT-BIH ECG databases. Our results show that the proposed FS algorithm outperforms the recursive feature elimination method in synthetic examples, and that the VF detector performance improves with the reduced feature set.

IMPLEMENTATION

MODULES:

- Dataset Collection
- Feature Selection
- Removing Irrelevant Features
- Booster accuracy

MODULES DESCSRIPTION:

Dataset Collection:

To collect and/or retrieve data about activities, results, context and other factors. It is important to consider the type of information it want to gather from your participants and the ways you will analyze that information. The data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable. after collecting the data to store the Database.

Feature Selection:

This is a hybrid measure of the prediction accuracy of the classifier and the stability of the selected features. Then the paper proposes Booster on the selection of feature subset from a given FS algorithm. The data sets from original data set by re sampling on sample space. Then FS algorithm is applied to each of these re sampled data sets to obtain different feature subsets. The union of these selected subsets will be the feature subset obtained by the Booster of FS algorithm. the Booster of an algorithm boosts not only the value of Q-statistic but also the prediction accuracy of the classifier applied. Several studies based on re sampling technique have been done to generate different data sets for classification problem , and some of the studies utilize re sampling on the feature space . The purposes of all these

studies are on the prediction accuracy of classification without consideration on the stability of the selected feature subset. FS algorithms— FAST, FCBF, and mRMR—and their corresponding Boosters, we apply k-fold cross validation. For this, k training sets and their corresponding k test sets are generated. For each training set, Booster is applied to obtain V . Classification is performed based on the training set with the selection V , and the test set is used for prediction accuracy

Removing Irrelevant Features:

The features of high dimensional microarray data are irrelevant to the target feature and the proportion of relevant features or the percentage of up-regulated Finding relevant features simplifies learning process and increases prediction accuracy. The finding, however, should be relatively robust to the variations in training data, especially in biomedical study, since domain experts will invest considerable time and efforts on this small set of selected features. The pre-processing steps to find weakly relevant features based on t-test and to remove irrelevant features based on MI.FS in high dimensional data needs preprocessing process to select only relevant features or to filter out irrelevant features. the selected subsets V1; . . . ; Vb obtained by s consist only of the relevant features where redundancies are removed, V will include more relevant features where redundancies are removed. Hence, V will induce smaller error of selecting irrelevant features. However, if s does not completely remove redundancies, V may result in the accumulation of larger size of redundant features. find more relevant features but may include more irrelevant features, and also may induce more redundant features. This is because no FS algorithm can select all relevant features while removing all irrelevant features and redundant features.

Booster accuracy:

The Booster of an FS algorithm that boosts the value of the Q-statistic of the algorithm applied. Empirical studies based on synthetic data Empirical studies show that the Booster of an algorithm boosts not only the value of Q-statistic but also the prediction accuracy of the classifier applied. Booster is simply a union of feature subsets obtained by a resembling technique. The resembling is done on the sample space. Booster needs an FS algorithm s and the number of partitions b. When s and b are needed to be specified, we will use notation s-Booster. Hence, s-Booster1 is equal to s since no partitioning is done in this case and the whole data is used. When s selects relevant features while removing redundancies, s-Booster will also select relevant features while removing redundancies. the notation FAST-Booster, FCBF-Booster, and mRMR-Booster for the Booster of the corresponding FS algorithm. we will evaluate the relative performance efficiency of s-Booster over the original FS algorithm s based on the prediction accuracy and Q-statistic.two Boosters, FAST-Booster , FCBF-Booster and mRMR-Booster. mRMR- Booster improves accuracy considerably: overall average accuracy. One interesting point to note here is that mRMR-Booster is more efficient in boosting the accuracy .we can observe that FAST-Booster also improves accuracy, but not as high as mRMR

SYSTEM ANALYSIS

EXISTING SYSTEM:

- One often used approach is to first discretize the continuous features in the preprocessing step and use mutual information (MI) to select relevant features. This is because finding relevant features based on the discretized MI is relatively simple while finding relevant features directly from a huge number of the features with continuous values using the definition of relevancy is quite a formidable task.
- Several studies based on resampling technique have been done to generate different data sets for classification problem and some of the studies utilize resampling on the feature space.
- The purposes of all these studies are on the prediction accuracy of classification without consideration on the stability of the selected feature subset.

DISADVANTAGES OF EXISTING SYSTEM:

- Most of the successful FS algorithms in high dimensional problems have utilized forward selection method but not considered backward elimination method

since it is impractical to implement backward elimination process with huge number of features.

- A serious intrinsic problem with forward selection is, however, a flip in the decision of the initial feature may lead to a completely different feature subset and hence the stability of the selected feature set will be very low although the selection may yield very high accuracy.
- Devising an efficient method to obtain a more stable feature subset with high accuracy is a challenging area of research.

PROPOSED SYSTEM:

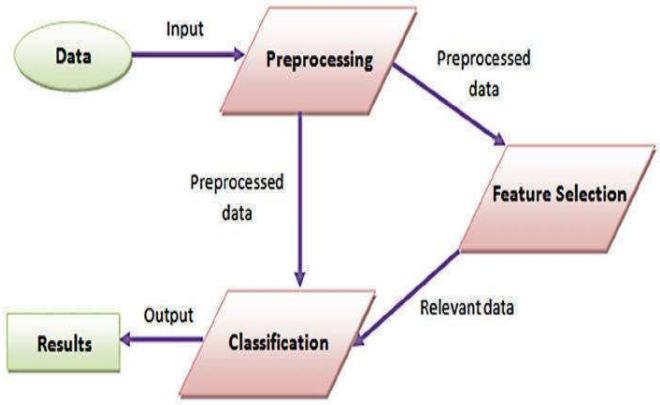
- This paper proposes Q-statistic to evaluate the performance of an FS algorithm with a classifier. This is a hybrid measure of the prediction accuracy of the classifier and the stability of the selected features. Then the paper proposes Booster on the selection of feature subset from a given FS algorithm.
- The basic idea of Booster is to obtain several data sets from original data set by resampling on sample space. Then FS algorithm is applied to each of these resampled data sets to obtain different feature subsets. The union of these selected subsets will be the feature subset obtained by the Booster of FS algorithm.

ADVANTAGES OF PROPOSED SYSTEM:

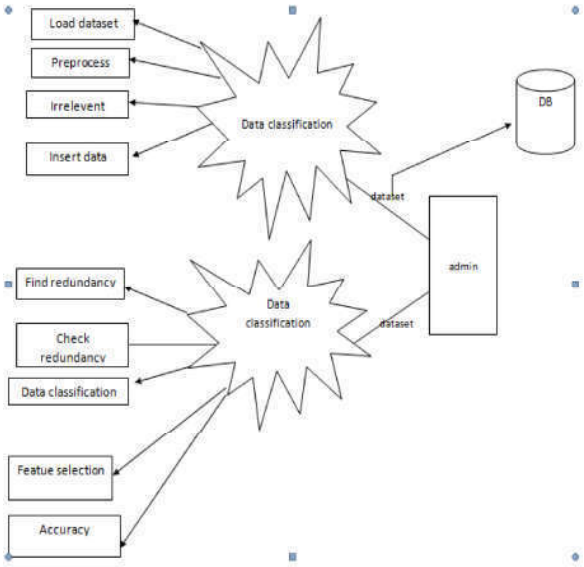
- Empirical studies show that the Booster of an algorithm boosts not only the value of Q-statistic but also the prediction accuracy of the classifier applied.
- We have noted that the classification methods applied to Booster do not have much impact on prediction accuracy and Q-statistic. Especially, the performance of mRMR-Booster was shown to be outstanding both in the improvements of prediction accuracy and Q-statistic.

SYSTEM DESIGN

SYSTEM ARCHITECTURE:



BLOCK DIAGRAM:

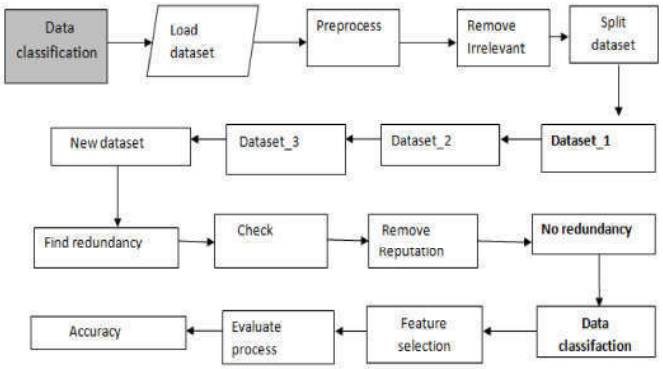


DATA FLOW DIAGRAM:

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts

information flow and the transformations that are applied as data moves from input to output.

4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

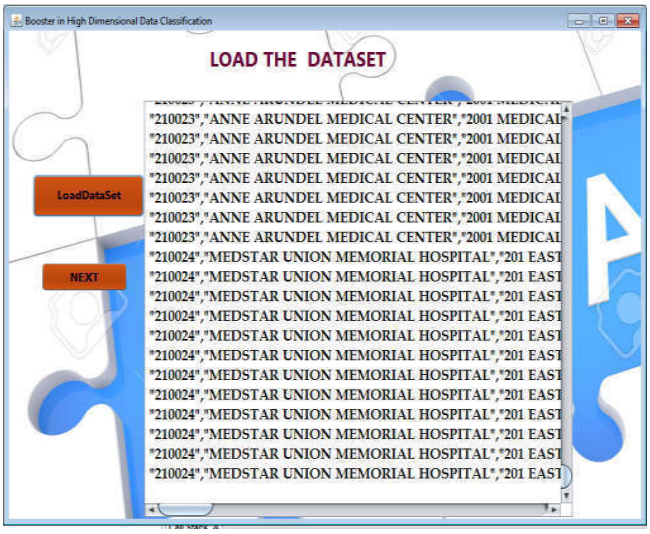
TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

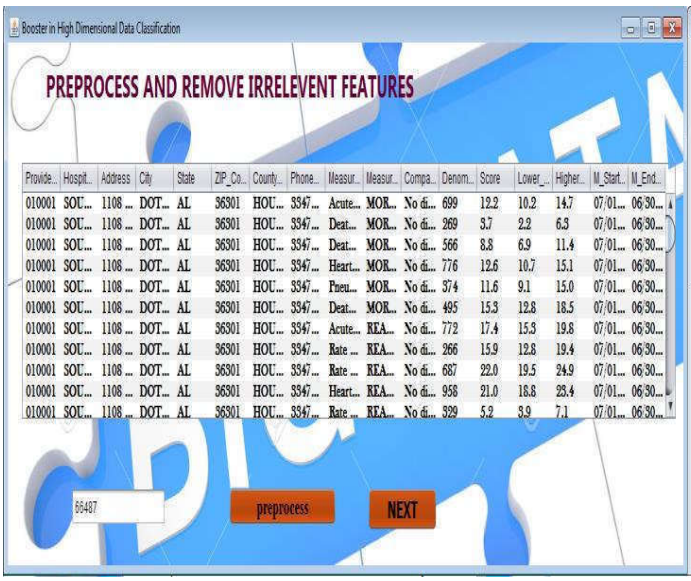
SOCIAL FEASIBILITY

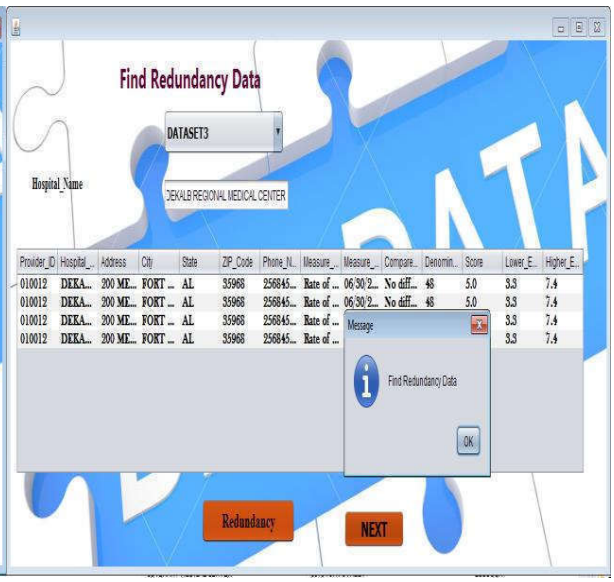
The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

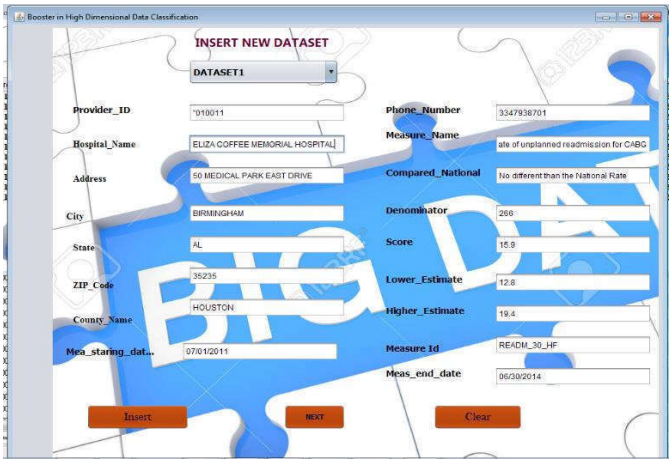
SCREEN SHOTS

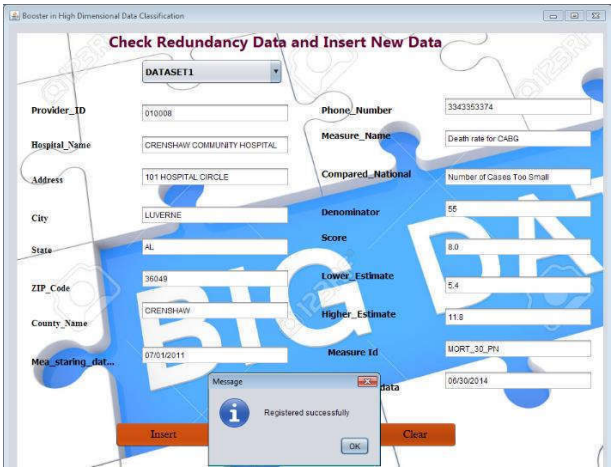


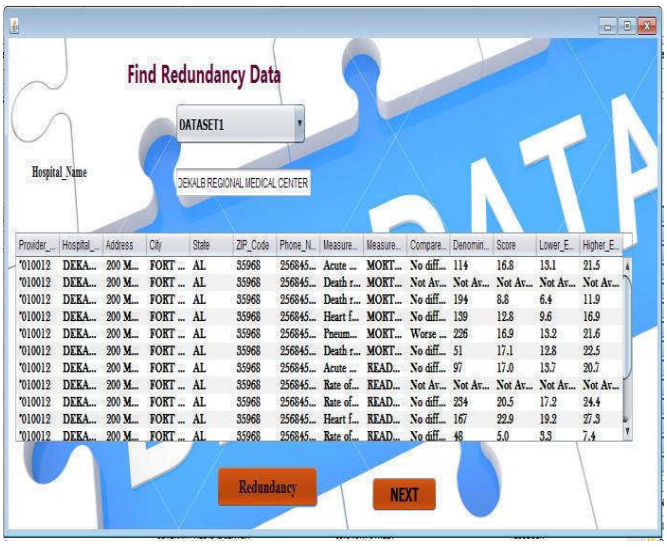
ISBN 978-81-939386-1-4

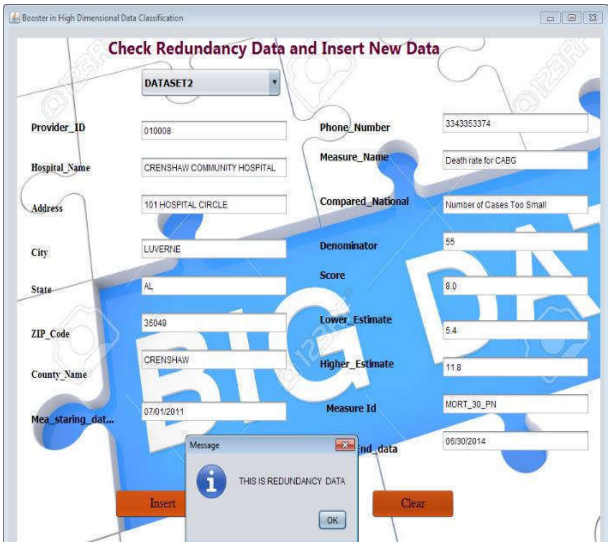




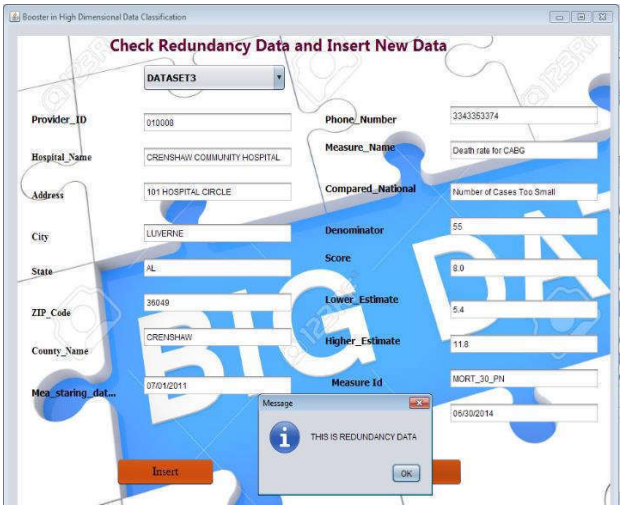




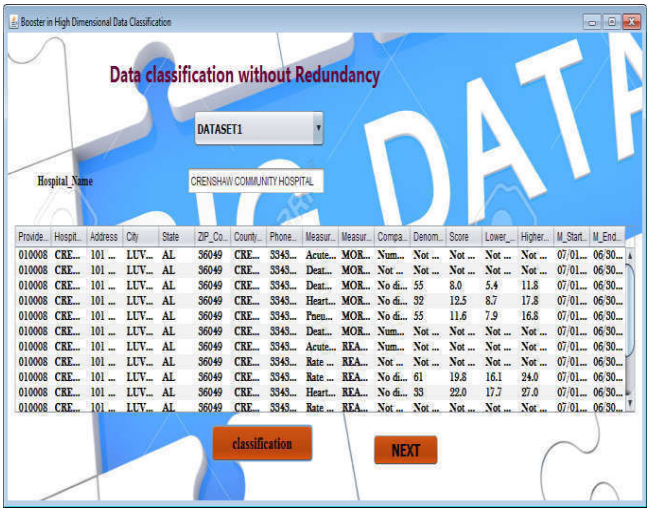


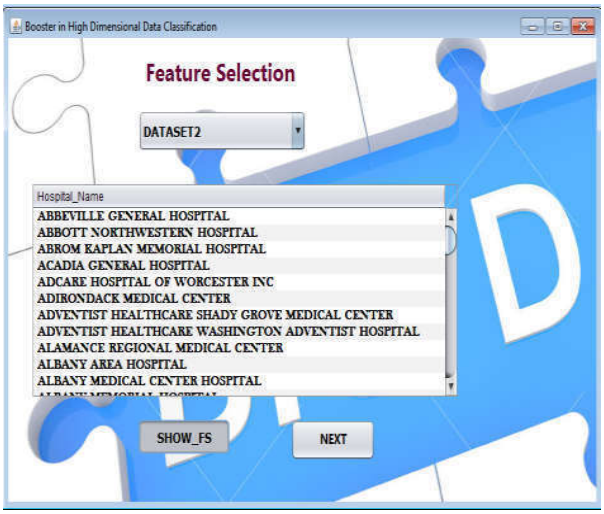


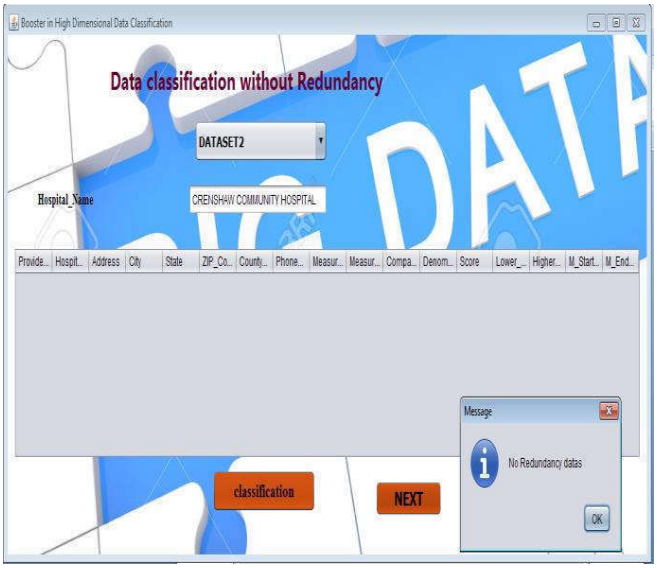
ISBN 978-81-939386-1-4

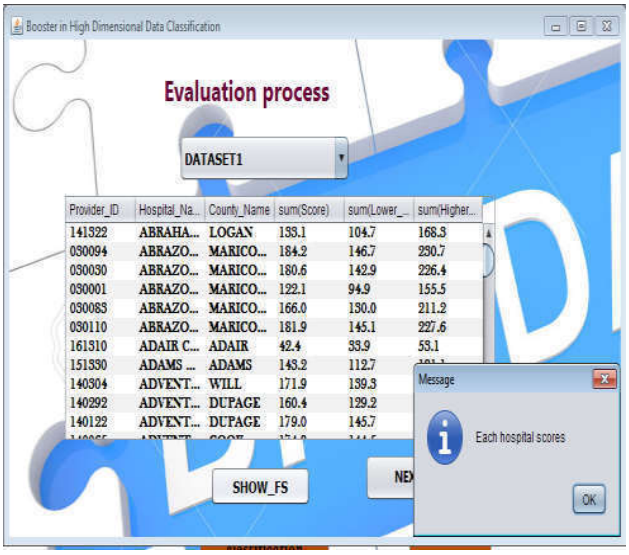


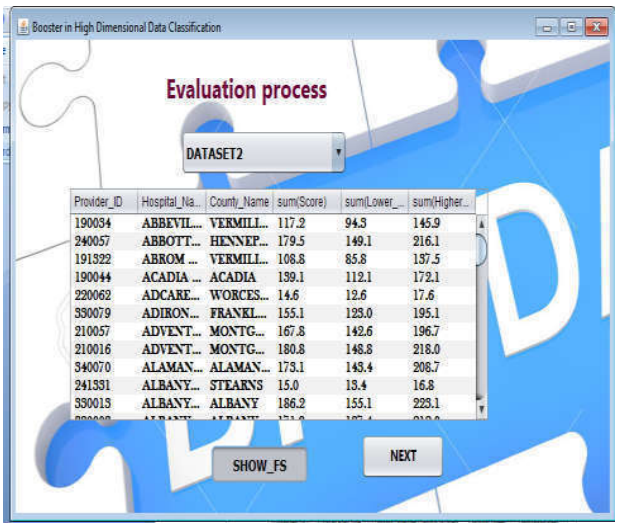


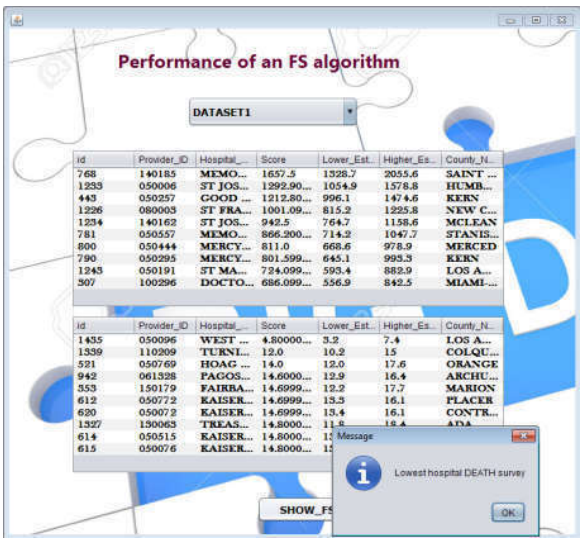


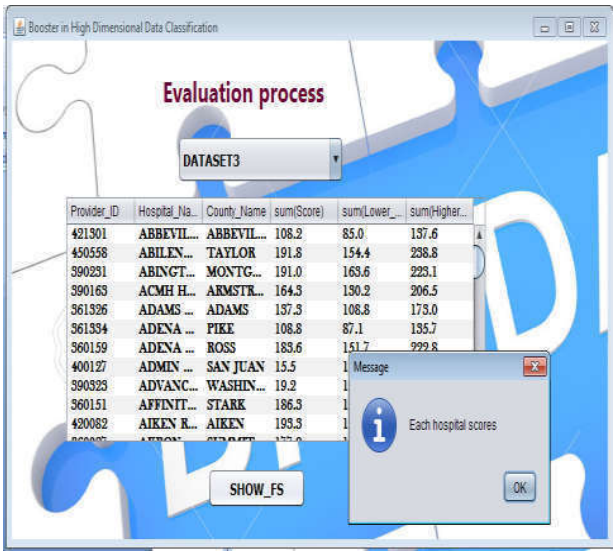


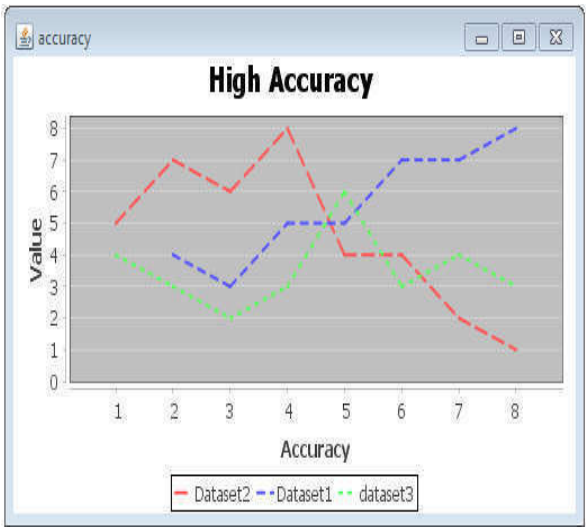




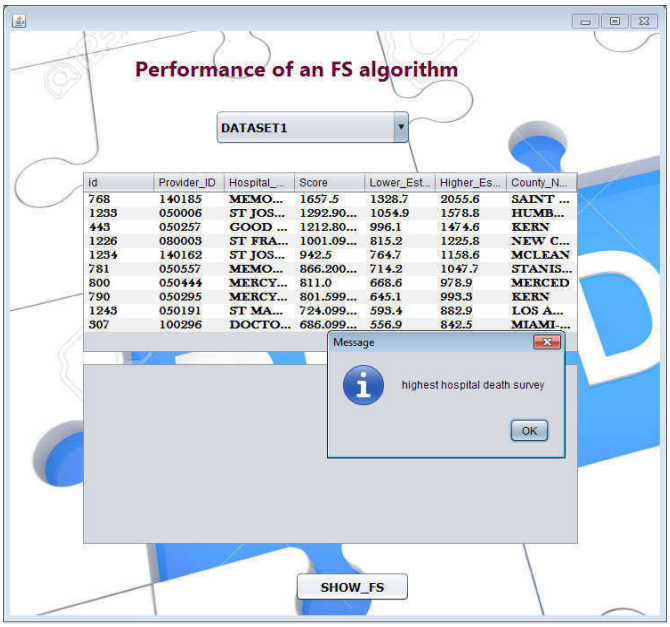








INPUT DESIGN AND OUTPUT DESIGN



INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

CONCLUSION

This paper proposed a measure Q-statistic that evaluates the performance of an FS algorithm. Q-statistic accounts both for the stability of selected feature subset and the prediction accuracy. The paper proposed Booster to boost the performance of an existing FS algorithm. Experimentation with synthetic data and 14 microarray data sets has shown that the suggested Booster improves the prediction accuracy and the Q-statistic of the three well-known FS algorithms: FAST, FCBF, and mRMR. Also we have noted that the classification methods applied to Booster do not have much impact on prediction accuracy and Q-statistic. Especially, the performance of mRMR-Booster was shown to be outstanding both in the improvements of prediction accuracy and Q-statistic.

It was observed that if an FS algorithm is efficient but could not obtain high performance in the accuracy or the Q-statistic for some specific data, Booster of the FS algorithm will boost the performance. However, if an FS algorithm itself is not efficient, Booster may not be able to obtain high performance. The performance of Booster depends on the performance of the FS algorithm applied. If Booster does not provide high performance, it implies two possibilities: the

data set is intrinsically difficult to predict or the FS algorithm applied is not efficient with the specific data set. Hence, Booster can also be used as a criterion to evaluate the performance of an FS algorithm or to evaluate the difficulty of a data set for classification.

REFERENCES

[1] T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," Bioinformatics, vol. 26, no. 3, pp. 392–398, 2010.

[2] D. Aha and D. Kibler, "Instance-based learning algorithms," Mach. Learn., vol. 6, no. 1, pp. 37–66, 1991.

[3] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. M. Izidore, S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. H. Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, vol. 403, no. 6769, pp. 503–511, 2000.

[4] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc. Nat. Acad. Sci., vol. 96, no. 12, pp. 6745–6750, 1999.

[5] F. Alonso-Atienza, J. L. Rojo-Alvare, A. Rosado-Mu~noz, J. J. Vinagre, A. Garcia-Alberola, and G. Camps-Valls, "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," Expert Syst. Appl., vol. 39, no. 2, pp. 1956–1967, 2012.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

An Overlay Architecture for Throughput Optimal Multipath Routing

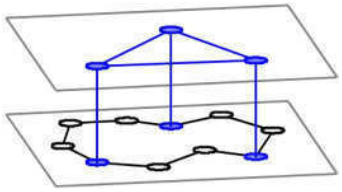
Sapa Mounika,
PG Student
Department of CSE,
Malla Reddy College of Engineering,
Hyderabad

Dr.J. Gladson Maria Britto,
Professor,
Department of CSE,
Malla Reddy College of Engineering,
Hyderabad

Abstract— **Legacy networks are often designed to operate with simple single-path routing, like the shortest path, which is known to be throughput suboptimal. On the other hand, previously proposed throughput optimal policies (i.e., backpressure) require every device in the network to make dynamic routing decisions. In this paper, we study an overlay architecture for dynamic routing, such that**
only a subset of devices (overlay nodes) need to make the dynamic routing decisions. We determine the essential collection of nodes that must bifurcate traffic for achieving the maximum multi-commodity network throughput. We apply our optimal node placement algorithm to severalgraphs and the results show that a small fraction of overlay nodes is sufficient for achieving maximum throughput. Finally, we a propose threshold-based policy (BP-T) and a heuristic policy (OBP), which dynamically control traffic bifurcations at overlay nodes. Policy BP-T is proved to maximize throughput for the case when underlay paths do no overlap. In all studied simulation scenarios, OBP not only achieves full throughput but also reduces delay in comparison to the throughput optimal back-pressure routing .

INTRODUCTION

An optimal routing in networks where some legacy nodes are replaced with overlay nodes. While the legacy nodes perform only forwarding on pre-specified paths, the overlay nodes are able to dynamically route packets. *Dynamic backpressure* is known to be an optimal routing policy, but it typically requires a homogeneous network, where all nodes participate in control decisions. Instead, we assume that only a subset of the nodes are controllable; these nodes form a network overlay within the legacy network.



A first finding is that ring networks require exactly 3 controllable (overlay) nodes to enable the same throughput region as when all nodes are controllable, independent of the total number of nodes in the network. Motivated by this, we develop an algorithm for choosing the minimum number of controllable nodes required to enable the full throughput region. We evaluate our algorithm on several classes of regular and random graphs. In the case of random networks with a power-law degree distribution, which is a common model for the Internet, we find that fewer than 80 out of 1000 nodes are required to be controllable to enable the full throughput region.

A. Motivation and Related Work

Backpressure (BP) routing, first proposed in, is a throughput optimal routing policy that has been studied for decades. Its strength lies in discovering multipath routes and utilizing them optimally without knowledge of the network parameters, such as arrival rates, link capacities, mobility, fading, etc. Nevertheless, the adoption of this

routing policy has not been embraced for general use on the Internet. This is due, in part, to an inability of backpressure routing to coexist with legacy routing protocols. With few exceptions, back-pressure routing has been studied in homogeneous networks, where all nodes are dynamically controllable and implement the backpressure policy across all nodes uniformly. As will be shown, backpressure routing as proposed in suboptimal when applied only to a subset of nodes in the network.

B.Problem Statement and Contributions

We consider two problem areas for control of heterogeneous networks. First, we develop algorithms for choosing the placement of controllable nodes, where our goal here is to allocate the minimum number of controllable nodes such that the full network stability region is available. Second, given any subset of nodes that are controllable, we also wish to develop an optimal routing policy that operates solely on these nodes.

THROUGHPUT REGION

The *throughput region* $\Lambda G(V)$ is the set of all arrival rates that can be achieved by any policy implemented at controllable nodes V on graph G. For the case where all nodes are controllable, i.e., $V = N$, the throughput region equals the stability region of graph G. This section characterizes this region for a given set of paths $P(V)$.

Packets destined for node c are called *commodity* c packets. Let $\lambda^c a$ be the rate of exogenous arrivals at node for commodity c, and let $\lambda = (\lambda^c a)$ be the multicommodity arrival rate vector for all sources a and commodities c.

Let $f_{ij}^{ab,c}$ be the *edge-flow* for commodity c on edge (i, j) along the shortest-path from node a to b. Flow for a path is allowed only

OPTIMAL PLACEMENT OF OVERLAY NODES

We would like to place controllable nodes to solve P1, but the constraint $\Lambda G(V) = \Lambda G(N)$ is difficult to evaluate directly. A simple implementation for P1 can use the fact that ΛG is a convex polytope, choosing the minimum number of controllable nodes to satisfy all points in the throughput region,The objectives of P1 and P2 are identical, therefore it is clear that P2 and P1 are equivalent. Thus, P2 provides a straightforward implementation of P1, however enumerating all extreme points in P2 may be impractical.

Instead of evaluating P2, we propose a surrogate condition that is easier to evaluate while still leading to the same optimal solution. Recall that the set of paths $P(V)$ includes all underlay paths P^{SP} and all n-concatenations (for any n) of these paths at controllable nodes V. Let PG be the set of all acyclic paths between all pairs of nodes in G. A first observation is that $P(N)$

This holds by the assumption that all 1-hop paths It included in the set P^{SP}, and since all nodes are controllable we can produce any path in G as a concatenation of 1-hop paths. Next, we define an important condition.

Condition C.1 (All-Paths): A set of controllable nodes V is said to satisfy the all-paths condition *if* $P(V) = PG$.

The condition requires the formation of all acyclic paths in a network. Since some of the paths are already given (in our paper P^{SP}), to satisfy the condition, a set of nodes V must enable all missing paths $PG \setminus P^{SP}$ by path concatenations. The following result establishes that this condition is necessary and sufficient for $\Lambda G(V) = \Lambda G$. In other words, to allow for maximum throughput achievability we must choose V to ensure that the path concatenations on these nodes form all missing paths in the network.

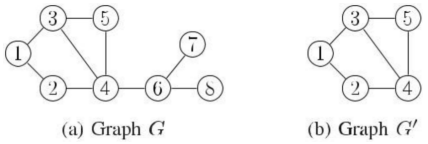
Theorem 1: Given a placement of controllable nodes V, satisfying the all-paths condition is necessary and sufficient for maximizing the throughput region, i.e.,

$$\Lambda G(V) = \Lambda G \text{ if and only if } P(V) = PG.$$

Overlay Node Placement Algorithm

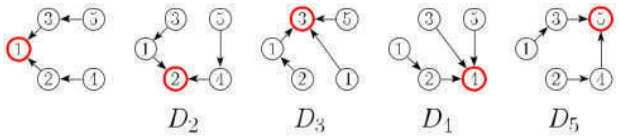
We design an algorithm to choose the placement of overlay nodes $V \subseteq N$ on a given graph $G = (N, E)$ such that the choice of overlay nodes is sufficient to satisfy the full throughput region of the network, i.e. $\Lambda G(V) = \Lambda G(N)$. At the end of this section we will show that the proposed algorithm optimally solves P3.

The algorithm consists of three phases: (1) removal of degree-1 nodes; (2) constraint pruning; and (3) overlay node placement. These phases are explained below, while each step is supported by a related claim which will help proving the optimality of the algorithm.

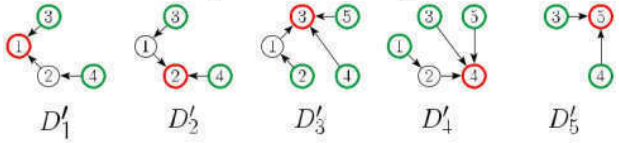


(a) Graph G (b) Graph G'

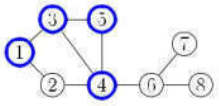
ISBN 978-81-939386-1-4



(c) Destination Trees D_n



(d) Pruned Trees D'_n



(e) Overlay Placement on G

While there exists any node $n \in N$ such that degree(n) = 1, set $N := N \setminus n$ and set $E := E \setminus e$, where e is the only edge that connects to node n. Repeat until no degree-1 nodes remain. All remaining nodes have a degree of at least 2, thus all attached trees have

The relationship between graphs G and G is shown for an example in Figs. 3a and 3b, where the subgraph connecting nodes 6, 7, and 8 formed an attached tree.

For any nodes $a, b \in N$, the proof considers four arrange-ments for the locations of a and b: (i) a and b are on the same attached tree, (ii) a is on some attached tree while b is not, (iii) b is on some attached tree while a is not.

OVERLAY NODES IN WIRELESS NETWORKS

The goal of this section is to motivate the need for additional study into the placement of overlay nodes for networks with wireless interference.

The all-paths condition C.1 is sufficient to achieve $\Lambda G(V) = \Lambda G$ in all networks, but this condition is not always a necessary condition in wireless networks. In other words, satisfying the all-paths condition may over allocate controllable nodes under certain wireless interference models. To see this, consider a clique where all edges have unit-capacity and all transmissions mutually interfere. Due to interference, the maximum network sum throughput in this scenario is one, and this maximum throughput can only be achieved when each source a sends to destination b directly over edge (a, b). Thus no multi-hop paths are required, and the all-paths condition is sufficient but not necessary for this scenario.

To illustrate an overlay network in a wireless scenario, we study the performance of the overlay node placement algorithm on random geometric graphs, which is a simple model for wireless networks with omnidirectional antennas. The geometric model has parameters N and r, where N is the number of nodes and r is the edge range. Random graphs are then generated by randomly placing N nodes in a unit square.

The results for random geometric graphs show that the overlay node placement algorithm chooses most nodes to be controllable. However, as noted above, the placement of con-trollable nodes by this algorithm is sufficient but may not be necessary for wireless networks. Thus, the minimum number of controllable nodes required to provide full throughput in wireless networks is unclear. A topic for future work is a study of the necessary conditions for $\Lambda G(V) = \Lambda G$ under various interference models.

CONCLUSION

We study optimal routing in legacy networks where only a subset of nodes can make dynamic routing decisions, while the legacy nodes can forward packets only on pre-specified shortest-paths. This model captures evolving heterogeneous networks where intelligence is introduced at a fraction of nodes.

We propose a necessary and sufficient condition for the overlay node placement to enable the full multicommodity throughput region. Based on this condition, we devise an algorithm for optimal controllable node placement. We run the algorithm on large random graphs to show that very often a small number of intelligent nodes suffices for full throughput .

Finally, we propose dynamic routing policies to be implemented in a network overlay. We provide a threshold-based policy that is optimal for overlays with non-overlapping tunnels, and provide and alternate policy for general networks that demonstrates superior performance in terms of both throughput and delay.

REFERENCES

•D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris, “Resilient overlay networks,” in *Proc. ACM SOSP*, Oct. 2001, pp. 131–145.

[Online]. Available: http://doi.acm.org/10.1145/502034.502048

- L. Bui, R. Srikant, and A. Stolyar, "Novel architectures and algorithms for delay reduction in back-pressure scheduling and routing," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 2936–2940.

- B. Fortz and M. Thorup, "Internet traffic engineering by opti-mizing OSPF weights," in *Proc. IEEE INFOCOM*, Mar. 2000.
-
- L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Found. Trends Netw.*, vol. 1, no. 1, pp. 1–144, 2006. [Online]. Available: http://dx.doi.org/10.1561/1300000001

- J. Han, D. Watson, and F. Jahanian, "Topology aware overlay networks," in *Proc. IEEE INFOCOM*, Mar. 2005, pp. 2554–2565. [Online]. Avail-able:

- N. M. Jones, "Practical algorithms for distributed network control," Ph.D. dissertation, Dept. Aeronautics Astronautics Massachusetts Inst. Technol., Cambridge, MA, USA, 2013.

- N. M. Jones, G. S. Paschos, B. Shrader, and E. Modiano, "An overlay architecture for throughput optimal multipath routing," in *Proc. ACM MobiHoc*, Aug. 2014, pp. 73–82.

- W. Khan, L. B. Le, and E. Modiano, "Autonomous routing algorithms for networks with wide-spread failures," in *Proc. IEEE MILCOM*, Oct. 2009, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5379792

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

SWEET: Serving the Web by Exploiting Email Tunnels

DEVISRAVANI
Department of Computer Science and Engineering
Mallareddy College of Engineering
Hyderabad, India

DR.P.JOHN PAUL
Department of computer science and Engineering
Mallareddy college of Engineering
Hyderbad,India

Abstract— **Open communications over the Internet pose serious threats to countries with repressive regimes, leading them to develop and deploy censorship mechanisms within their networks. Unfortunately, existing censorship circumvention systems do not provide high *availability* guarantees to their users, as censors can easily identify, hence disrupt, the traffic belonging to these systems using today's advanced censorship technologies. In this paper, we propose Serving the Web by Exploiting Email Tunnels (SWEET), a highly available censorship-resistant infrastructure. SWEET works by encapsulating a censored user's traffic inside email messages that are carried over public email services like Gmail and Yahoo Mail. As the operation of SWEET is not bound to any specific email provider, we argue that a censor will need to block email communications all together in order to disrupt SWEET, which is unlikely as email constitutes an important part of today's Internet. Through experiments with a prototype of our system, we find that SWEET's performance is sufficient for Web browsing. In particular, regular Websites are downloaded within couple of seconds.**

***Index Terms*— Censorship circumvention, email communications, traffic encapsulation.**

I. Introduction

The Internet provides users from around the world with an environment to freely communicate, exchange ideas and information. However, free communication continues to threaten repressive regimes, as the open circulation of information and speech among their citizens can pose serious threats to their existence. Recent unrest in the middle east demonstrates that the Internet can be widely used by citizens under these regimes as a very powerful tool to spread censored news and information, inspire dissent, and organize events and protests. As a result, repressive regimes extensively monitor their citizens' access to the Internet and restrict open access to public networks by using different technologies, ranging from simple IP address blocking and DNS hijacking to the more complicated and resource-intensive Deep Packet Inspection (DPI).

With the use of censorship technologies, a number of different systems were developed to retain the openness of the Internet for the users living under repressive Regimes. The earliest circumvention tools are HTTP proxies that simply intercept and manipulate a client's HTTP requests, defeating IP address blocking and DNS hijacking techniques. The use of more advanced censorship technologies such as DPI, rendered the use of HTTP proxies ineffective for circumvention. This led to the advent of more advanced tools such as Ultra surf and Psiphon, designed to evade content filtering. While these circumvention tools have helped, they face several challenges. We believe that the biggest one is their lack of *availability*, meaning that a censor can disrupt their service frequently or even disable them completely. The common reason is that the network traffic made by these systems can be distinguished from regular Internet traffic by censors, i.e., such systems are not unobservable For example, the popular Tor network works by having users connect to an ensemble of nodes with public IP addresses, which proxy users' traffic to the requested, censored destinations. This public knowledge about Tor's IP addresses, which is required to make Tor usable by users globally, can be and is being used by censors to block their citizens from accessing Tor. To improve availability, recent proposals for circumvention aim to make their traffic unobservable to the censors by pre-sharing secrets with their clients. Others suggest to conceal circumvention by making infrastructure modifications to the Internet. Nevertheless, deploying and scaling these systems is a challenging problem, as discussed in Section II.

A more recent approach in designing unobservable circumvention systems is to imitate popular applications like Skype and HTTP, as suggested by Skype-Morph, Censor Spoofer, and StegoTorus. However, it has recently been shown [29] that these systems' unobservability is breakable; this is because a comprehensive imitation of today's complex protocols is sophisticated and infeasible in many cases. A promising alternative suggested, is to not mimic protocols, but run the actual protocols and find clever ways to tunnel the hidden content into their genuine traffic; this is the main motivation of the approach taken in this paper.

In this paper, we design and implement SWEET, a censorship circumvention system that provides high availability by leveraging the openness of email communications. Fig. 1 shows the main architecture. A SWEET client, confined by a censoring ISP, tunnels its network traffic inside a series of email messages that are exchanged between herself and an email server operated by SWEET's server. The SWEET server acts as an Internet prox

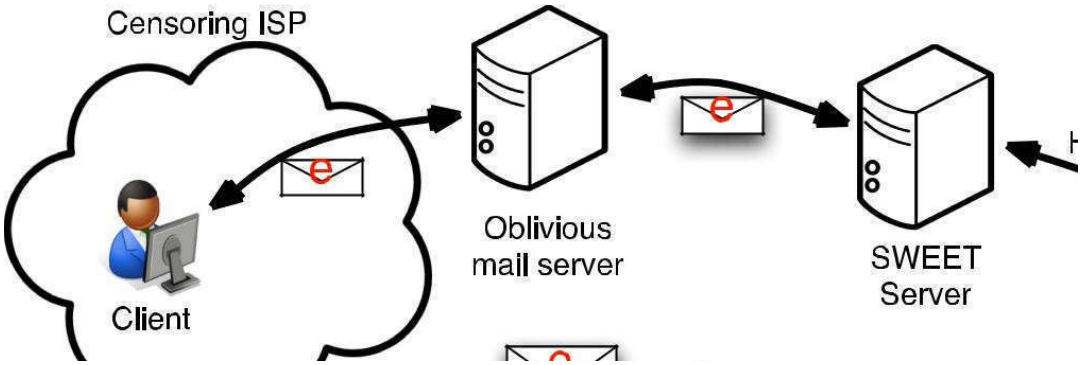


Fig. 1. Overall architecture of SWEET.

traffic to the requested blocked destinations. The SWEET client uses an oblivious, public mail provider (e.g., Gmail, Hotmail, etc.) to exchange the encapsulating emails, rendering standard email filtering mechanisms ineffective in identifying/ blocking SWEET-related emails. More specifically, to use SWEET for circumvention a client needs to create an email account with *some* public email provider; she also needs to obtain SWEET's client software from an out-of-bound channel (similar to other circumvention systems). The user configures the installed SWEET software to use her public email account, which sends/receives encapsulating emails on behalf of the user to/from the email address of SWEET.

SWEET's unobservability: We claim that a censor is not easily able to distinguish between SWEET's email messages and benign email messages. As described later in Section IV, a SWEET client has two options in choosing her email account:
1) *Alien Mail* a non-domestic email that encrypts emails (e.g., Gmail for users in China), and 2) *Domestic Mail* a domestic email account with no need for encryption (e.g., 163.com for users in China). As described in Section IV, when Alien Mail is used by a client all of its SWEET emails are sent to a publicly known email address, e.g., `tunnel@sweet.org`, encrypted; however, a censor will not be able to identify these emails since they are *proxied* by the Alien Mail server running outside the censoring area. In simpler words, the censor only observes that the client is exchanging encrypted messages with the Alien Mail server (e.g., Gmail's mail server in U.S.), but he will not be able to observe neither the recipient's email address (`tunnel@sweet.org`), nor the IP address of the sweet.org mail server. As a result, existing approaches for spam filtering such as shooting the spamming SMTP servers and dropping spam emails are entirely infeasible. In the case of Domestic Mail, the SWEET server uses a secondary *secret* email account, which is only shared with that particular client, for exchanging SWEET emails (i.e., `myotheremail@163.com` instead of `tunnel@sweet.org` address). As a result, the censor will not be able to identify SWEET messages from their recipient fields (since the censor does not know the association of `myotheremail@163.com` with SWEET). Also, the use of steganography/encryption to embed tunneled data renders DPI infeasible.

SWEET's availability: Given SWEET's unobservability dis-cussed above, a censor can not efficiently distinguish between SWEET emails and benign email messages. Hence, in order to block SWEET a censor needs to block all email messages to the outside world. However, email is an essential service in today's Internet and it is very unlikely that a censorship authority will block *all* email communications to the outside world, due to different financial and political reasons. This, along the fact that SWEET can be reached through a wide range of domestic/non-domestic email providers provides a high degree of *availability* for SWEET.

Prototype implementation: We have built a prototype implementation for SWEET and evaluated its performance. We have also proposed and prototyped two different designs for SWEET client. The first client design uses email protocols, e.g., POP3 and SMTP, to communicate with the SWEET system, and our second design is based on using the webmail interface. Our measurements show that a SWEET client is able to browse regular-sized web destinations with download times in the order of couple of seconds.

In fact, the high availability of SWEET comes for the price of higher, but bearable, communication latencies. Fig. 2 compares SWEET with several popular circumvention systems regarding their availability and communication latency. As our measurements in Section VII show, SWEET provides communication latencies that are convenient for latency-sensitive activities like web browsing (i.e., few seconds). Such additional, tolerable latency of SWEET comes with the bonus of better availability, as discussed in Section V-B.

Our contributions: In summary, this paper makes the following main contributions: i) we propose a novel infrastructure for censorship circumvention, SWEET, which provides high avail-ability, a feature missing in existing circumvention systems;
ii) we develop two prototype implementations for SWEET (one using webmail and the other using email exchange protocols) that allow the use of nearly all email providers by SWEET clients; and, iii) we show the feasibility of SWEET for practical censorship circumvention by measuring the communication latency of SWEET for web browsing using our prototype implementation.

Paper's organization: The rest of this paper is organized as follows; in Section II, we discuss the related work on

ISBN 978-81-939386-1-4

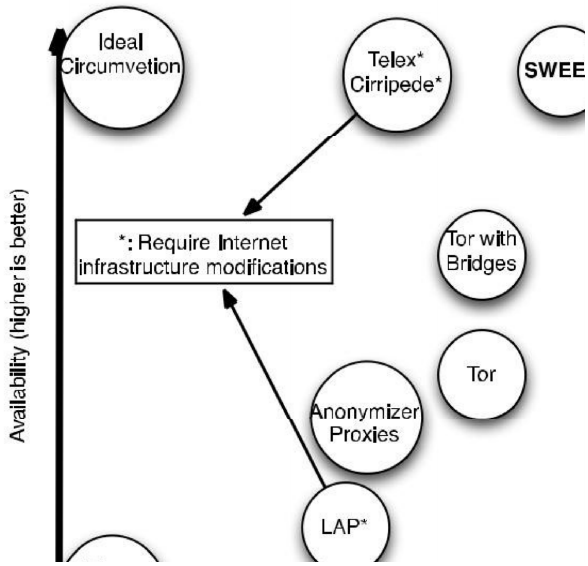


Fig. 2. Availability and communication latency comparison of circumvention systems.

unobservable censorship circumvention. In Section III, we reviews our threat model. We provide the detailed description of the proposed circumvention system, SWEET, in Section IV. We discuss SWEET's censorship features, including its avail-ability, in Section V and compare it with the literature. Our prototype implementation and evaluations are presented in Sections VI and VII, respectively. Finally, we conclude the paper in Section VIII.

II. Related Work

There has been much work on unobservable censorship circumvention systems Similar to SWEET, Free Wave , Cloud Transport, and Covert Cast also work by tunneling circumvention traffic into the actual runs of popular network protocols. For instance, Free Wave tunnels Internet traffic inside VoIP communications. This tunneling approach provides much stronger unobservability against the censors compared to imitation-based circumvention systems, as demonstrated by Houmansadr

Several designs seek unobservability by sharing secret information with their clients, which are not known to censors. For instance, the Tor network has recently adopted the use of *Tor Bridges*, a set of volunteer nodes connecting clients to the Tor network, whose IP addresses are selectively distributed among Tor users by Tor. As another example, In franet shares a secret key and some secret URL addresses with a client, which is then used to establish an unobservable communication between the client and the system. Collage works by having a client and the system *secretly* agree on some user-generated content sharing websites, e.g., flickr.com, and communicate using steganography. Unfortunately, sharing secrets with a wide range of clients is a serious challenge, as a censor can obtain the same secret information by pretending to be a client.

Some recent research suggests circumvention being built into the Internet infrastructure to better provide unobservability. These systems rely on collaboration from some Internet routers that intercept users' traffic to uncensored destinations to establish covert communication between the users and the censored destinations. Telex and Cirripede provide this unobservable communication without the need for some pre-shared secret information with the client, as the secret keys are also covertly communicated inside the network traffic. Cirripede uses an additional client registration stage that provides some advantages and limitations as compared to Telex and Decoy routing systems. Recent studies investigate the real-world deployment of decoy routing systems by evaluating the placement of decoy routers on the Internet in adversarial settings.

There are two projects that work in a similar manner to SWEET: FOE and Mail My Web . Instead of tunneling traffic, which is the case in SWEET, these systems simply download a requested website and send it as an email attachment to the requesting user. This highly limits their performance compared to SWEET, as discussed in Section IV-D.

III. Threat Model

We assume that a user is confined inside a censoring ISP. The ISP blocks the user's access to certain Internet destinations, namely *blocked destinations*. The censor is assumed to use today's advanced filtering technologies, including IP address blocking, DNS hijacking, and deep packet inspection technique. The ISP also monitors all of its egress/ingress traffic to detect any use of circumvention techniques.

We assume that the censorship is constrained not to degrade the *usability* of the Internet. In other words, even though it *selectively* blocks certain Internet connections, she is not willing to block key Internet services *entirely*. In particular, the operation of SWEET system relies on the fact that a censoring ISP does not block *all* email communications, even though she can selectively block emails/email providers. We also assume that the ISP has as much information about SWEET as any SWEET client.

We also consider active behaviors of the ISP. In addition to traffic monitoring, the censor manipulates its Internet traffic, e.g., by selectively dropping packets, and adding latency to some packets, to disrupt the use of circumvention systems and/or to detect the users of such systems. Again, such perturbations are constrained to preserve the usability of the Internet for benign users.

IV. Design of SWEET

In this section, we describe the detailed design of SWEET. Fig. 1 shows the overall architecture. SWEET tunnels network connections between a client and a server, called SWEET server, inside email communications. Upon receiving the tunneled network packets, the SWEET server acts as a transparent proxy between the client and the network destinations requested by the client.

ISBN 978-81-939386-1-4

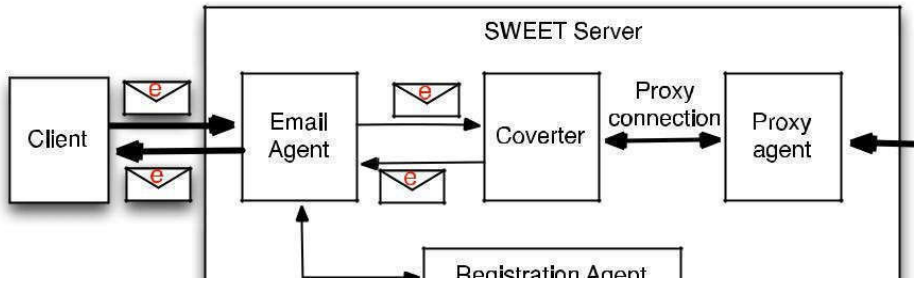


Fig. 3. The main architecture of SWEET server.

A client's choices of email services: A SWEET client has two options for his email provider: *Alien Mail*, and *Domestic Mail*.

1) **Alien Mail** An Alien Mail is a mail provider whose mail servers reside outside the censoring ISP, e.g., Gmail for the Chinese clients. We only consider Alien Mails that provide email encryption, e.g., Gmail and Hush mail. A SWEET client who uses an Alien Mail does not need to apply any additional encryption/steganography to her encapsulated contents. Also, she simply sends her emails to the publicly advertised email address of SWEET server, e.g., `tunnel@sweet.org`, since the censors will not be able to observe (and block) the `tunnel@sweet.org` address inside SWEET messages, which are exchanged between the client and the Alien Mail server in an encrypted format.
2) **Domestic Mail** A Domestic Mail is an email provider hosted inside the censoring ISP and possibly collaborating with the censors, e.g., 163.com for the Chinese clients. Since the censors are able to observe the email contents, the SWEET client using a Domestic Mail should hide the encapsulated contents through steganography (e., by doing image/text steganography inside email messages). Also, the client cannot send her SWEET emails to the public email address of SWEET server (`tunnel@sweet.org`) since the mail recipient field is observable to the Domestic Mail provider and/or the censor. Instead, the client generates a secondary email address, `myotheremail@somedomain.com` (which could be either Domestic Mail or Alien Mail), and then provides the email credentials for this secondary account *only* to SWEET server through an out-of-band channel (e.g., through an online social network). The SWEET server uses this email address to exchange SWEET emails *only* with this particular client.

In the following, we describe the details of SWEET's server and client architectures. To avoid confusion and without loss of generality, we only consider the case of Alien Mail being used by the client. If Domestic Mail is used, the client and server should also perform some steganography operations to hide the encapsulated traffic, as well as they should exchange a secondary email address, as described above.

A. SWEET Server

The SWEET server is the part of SWEET running outside the censoring region. It helps SWEET clients to evade censorship by proxying their traffic to blocked destinations. More specifically, a SWEET server communicates with censored users by exchanging emails that carry tunneled network packets. Fig. 3 shows the main design of SWEET server, which is composed of the following elements:

① *Email agent:* The email agent is an IMAP and SMTP server that receives emails that contain the tunneled Internet traffic, sent by SWEET clients to SWEET's email address. The email agent passes the received emails to another components of the SWEET server, the converter and the registration agent. The email agent also sends emails to SWEET clients, which are generated by other components of SWEET server and contain tunneled network packets or client registration information.

② *Converter:* The converter processes the emails passed by the email agent, and extracts the tunneled network packets. It then forwards the extracted data to another component, the proxy agent. Also, the converter receives network packets from the proxy agent and converts them into emails that are targeted to the email address of corresponding clients. The converter then passes these emails to the email agent for delivery to their intended recipients. As described later, the converter encrypts/decrypts the email attachments of a user using a secret key shared with that user.

③ *Proxy agent:* The proxy agent proxies the network packets of clients that are extracted by the converter, and sends them to the Internet destination requested by the clients. It also sends packets from the destination back to the converter.

④ *Registration agent:* This component is in charge of registering the email addresses of the SWEET clients, prior to their use of SWEET. The information about the registered clients can be used to ensure quality of service and to prevent denial-of-service attacks on the server. Additionally, the registration agent shares a secret key with the client, which is used to encrypt the tunneled information between the client and the server.

The email agent of the SWEET server receives two type of emails; *traffic emails*, which contain tunneled traffic from the clients (sent to `tunnel@sweet.org`), and

registration emails, which carry client registration information

Client registration: Before the very first use of the SWEET service, a client needs to register her email address with the system. This is automatically performed by the client's SWEET software. The objective of client registration is twofold: to prevent denial-of-service (DoS) attacks and to share a secret key between a client and the server. A DoS attack might be launched on the server to disrupt its availability, e.g., through sending many malformed emails on behalf of non-existing email addresses (this is discussed in Section V). In order to register (or update) the email address of a client, the client's SWEET software sends a registration email from the user's email address, to the SWEET's registration email address. i.e., `register@sweet.org`, requesting registration. The email agent forwards all received registration emails to the registration agent (④). For any new registration request, the registration agent generates and sends an email to the requesting email address (through the email agent) that contains a unique computational *challenge* (e.g., [41]). After solving the challenge, the client software sends a second email to `register@sweet.org` that contains the solution to the challenge, along with a Diffie-Hellman [42] public key $K_C = g^{k_C}$. If the client's response is verified by the registration agent, the client's email address will be added to a *registration list*, that contains the list of registered email addresses with their expiration time. Also, the registration agent uses its own Diffie-Hellman public key, $K_R = g^{k_R}$, to evaluate a shared key

$k_{C,R} = g^{k_R k_C}$

for the later communications with the client. The registration agent adds this key to the client's entry in the registration list, to be used for communications with that client. The client is able to generate the same $k_{C,R}$ key using SWEET's publicly advertised public key and her own private key [42].

Tunneling the traffic: Any traffic email received by the email agent is processed as follows: the email agent (①) forwards the email to the converter(②). The converter processes the traffic email and extracts the tunneled information. The converter, then, decrypts the extracted information (using the key $k_{C,R}$ corresponding to the user) and sends it to the proxy agent (③). Finally, the proxy processes the received packet as required, e.g., sends the packet to the requested destination. Simi-larly, for any tunneled packet received from the proxied destinations, the proxy agent sends it to the converter. The converter encrypts the received packet(s) (using the corre-sponding $k_{C,R}$), and generate a traffic email containing the encrypted data as an attachment, targeted to the email address of the corresponding client. The generated email is passed to the email agent, who sends the email to the client. Note that to improve the latency performance, small packets that arrive at the same time get attached to the same email.

B. SWEET Client

To use SWEET, a client needs to obtain a copy of SWEET's client software and install it on her machine. The client also needs to create one or two email account (depending on if she uses an Alien Mail or a Domestic Mail for her primary email). A client needs to configure the installed SWEET's software with information about her email account. Prior to the first use of SWEET by a client, the client software registers the email address of its user with the SWEET server and obtains a shared secret key $k_{C,R}$, as described in Section IV-A.

We propose two designs for SWEET client: a protocol-based design, which uses standard email protocols to exchange email with client's email provider, and a webmail-based design, which uses the webmail interface of the email provider. We describe these two designs in the following.

1) Protocol-Based Design: Fig. 4(a) shows the three main elements.

❶ *Web Browser:* The client can use any web browser that supports proxying of connections, e.g., Google Chrome, Internet Explorer, or Mozilla Firefox. The client needs to con-figure her browser to use a local proxy server, e.g., by setting `localhost:4444` as the HTTP/SOCKS proxy. The client can use two different browsers for browsing with and without SWEET to avoid the need for frequent re-configurations of the browser. Alternatively, some browsers (e.g., Chrome, and Mozilla Firefox) allow a user to have multiple browsing *profile*s, hence, a user can setup two profiles for browsing with and without SWEET.

❷ *Email Agent:* It sends and receives SWEET emails thor-ough the client's email account. The client needs to configure it with the settings of the SMTP and IMAP/POP3 servers of her email account. The client also needs to provide it with the account login information.

❸ *Converter:* It sits between the web browser and the email agent, and converts SWEET emails into network packets and vice versa. It uses the keys shared with SWEET, $k_{C,R}$, to encrypt/decrypt email content.

Once the client enters a URL into the configured browser (❶), the browser makes a proxy connection to the local port that the converter (❸) is listening on. The converter accepts the proxy connection and keeps the state of the estab-lished TCP/IP connections. For packets that are received from the browser, the converter generates traffic emails, targeted to `tunnel@sweet.org`, having the received packets as encrypted email attachments (using the key $k_{C,R}$). Such emails are passed to the email agent (❷) that sends the emails to the SWEET server through the public email provider of the client.

The email agent is also configured to receive emails from the client's email account through an email retrieval pro-tocol, e.g., IMAP or POP3. This allows the email agent to continuously look for new emails from the server. Once new emails are received, the email agent passes them to the converter, who in turn extracts the packets from the emails, decrypts them, and sends them to the browser over the existing TCP/IP connection.

2) Webmail-Based Design: Alternatively, the SWEET client can use the webmail interface of the client's public email provider. as showed in Fig. 4(b). The main difference with the protocol-based design is that in this case the email agent (❷) uses a web browser to exchange emails. More specifically, the email agent uses its web browser to open a webmail interface with the client's email account, using the user's authentication

ISBN 978-81-939386-1-4

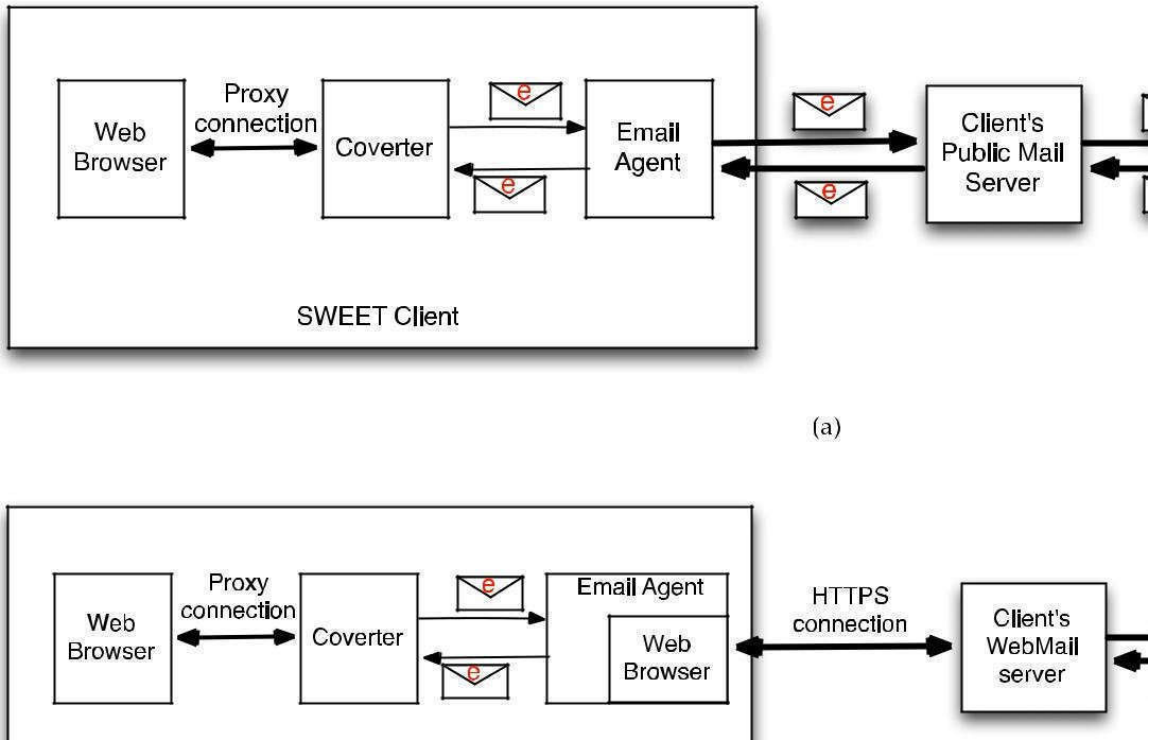


Fig. 4. Design of SWEET client software. (a) The protocol-based design. (b) The webmail-based design.

credentials for logging in. Through this HTTP/HTTPS connection, the email agent communicates with the SWEET server by sending and receiving emails.

C. The Choice of the Proxy Protocol

As mentioned before, the SWEET server uses a proxy agent to receive the tunneled traffic of clients and to establish connections to the requested destinations. We consider the use of both SOCKS [31] and HTTP [43] proxies in the design, as each provides unique advantages. Our server's proxy agent runs a SOCKS proxy and an HTTP proxy in parallel, each on a different port. A user can choose to use the type of proxy by configuring her client to connect to the corresponding port.

The use of the SOCKS proxy allows the client to make *any* IP connection through the SWEET system, including dynamic web communications, such as Java script or AJAX, and instant messaging. In contrast, an HTTP proxy only allows access to HTTP destinations. However, an HTTP proxy may speed up connections by using HTTP-layer optimizations such as caching or pre-fetching of web objects.

D. An Alternative Approach: Web Download

A trivial approach in providing censorship circumvention using email is to download an entire webpage and attach it as an email attachment to emails that are targeted to the requesting users. In fact, this approach is under development by the open-source foe project, nd the for-profit service of Mail My Web. Unfortunately, this simple approach only provides a limited access to the Internet: a user can only access static websites. In particular, this approach cannot be used to access destinations that require end-to-end encryption, contain dynamic web applications like HTML5 and Javascript sockets, or need user login information. Also, this approach does not support accessing web destinations that require a live Internet connection, e.g., video streaming websites, instant messaging, etc. In fact, the MailMyWeb service uses some heuristics to tackle some of these shortcomings partially, which are privacy-invasive and inefficient. For example, in order to access login-based websites MailMyWeb requests a user to send her login credentials to MailMyWeb by email. Also, a user can request for videos hosted *only* on the YouTube video sharing website, which are then downloaded by MailMyWeb and sent as email attachments; this causes a large delay between the time a video is requested until it is has received by the user. SWEET, on the other hand, provides a comprehensive web browsing experience to its users since it can tunnel any kind of IP traffic.

V. Discussions and Comparisons

In this section, we evaluate SWEET's circumvention capabilities by discussing important features that are essential for an effective circumvention.

A. Unobservability

We say a circumvention tool provides unobservability if censors are not able to identify neither the traffic, nor the clients using that tool. Unobservability has been considered as a main feature in the design of recent circumvention systems [22]–[24], [26]–[28].

We claim that a censor is not easily able to distinguish between SWEET's email messages and benign email messages. As described in Section IV, a SWEET client has two options in choosing her email account: 1) *Alien Mail* a non-domestic email that encrypts emails (e.g., Gmail for users in China), and 2) *Domestic Mail* a domestic email account with no need for encryption (e.g., 163.com for the users in China). When Alien Mail is used by a SWEET client all of its SWEET emails are encrypted and are exchanged with a publicly known email address of SWEET, e.g., `tunnel@sweet.org`; however, a censor will not be able to identify these SWEET emails since they are *proxied* by the Alien Mail server running outside the censoring area. In simpler words, the censor only observes that the client is sending/receiving messages with an Alien Mail server (e.g., Gmail's mail server in U.S.), but he will not be able to observe neither the recipient's email address (`tunnel@sweet.org`), nor the IP address of the sweet.org mail server. As a result, existing approaches for spam filtering such as shooting spamming SMTP servers and dropping spam emails are entirely infeasible. In the case of Domestic Mail, the SWEET server uses a *secret* secondary email account, which is only shared with that particular client, for exchanging SWEET emails (i.e., `myotheremail@163.com` instead of `tunnel@sweet.org` address). As a result, the censor will not be able to identify SWEET messages from their recipient field (since the censor can not associate the private email address with SWEET). Also, the use of steganography to embed tunneled data renders DPI infeasible.

In addition, to ensure unobservability the user's email traffic patterns should mimic that of normal email communications, to defeat traffic analysis by a censor; this limits the bandwidth available to the user, as discussed in Section VII-B.

B. Availability

SWEET's availability is tied to the assumption that a censor is not willing to block *all* email communications. As the use of SWEET does not require using any specific email provider, users can always find an email service to get connected to SWEET. IP filtering and DNS hijacking would not be able to stop SWEET traffic as a SWEET user's traffic is destined to her public email provider, but not to an IP address or nameserver belonging to the SWEET system. Moreover, deep packet inspection (DPI) is rendered ineffective due to the use of encrypted emails in the case of Alien Mail, and steganography in the case of Domestic Mail.

As another approach to disrupt the operation of SWEET, a censor might try to launch a denial-of-service (DoS) attack on SWEET server. The common techniques for DoS attacks, e.g., ICMP flooding and SYN flooding, can be mitigated by protecting the SWEET server using up-to-date firewalls. Alternatively, a censor can play the role of a SWEET client and send traffic through its SWEET client software in a way that overloads the SWEET server. As an example, the attacker can flood the SWEET's SOCKS proxy by initiating many incomplete SOCKS connections, or sending SYN floods. A censor could even send such attacking requests on behalf of a number of rogue (non-existing) email addresses, to render an email blacklist maintained by SWEET server ineffective in preventing such attacks. To protect against possible DoS attacks, SWEET requires a new user to register her email address with SWEET server prior to her first use. Such registration can be performed in an unobservable manner by SWEET's client software through the email communication channel (see Section IV-A). Also, to ensure the quality of service for all users, the SWEET server can limit the use of SWEET by putting a cap on the volume of traffic communicated by each registered email address.

C. Other Properties of SWEET

Confidentiality: As mentioned before, SWEET encrypts the tunneled traffic, i.e., email attachments are encrypted using a key shared between a user and SWEET server. This ensures the confidentiality of user communications from any entity wiretapping the traffic, including the censorship authorities and the public email provider. Note that the email attachments are encrypted even if the user choose a plaintext email service. To make a connection confidential from SWEET server, the user can use an end-to-end encryption with the final destination, e.g., by using HTTPS, or alternatively the user can use SWEET to connect to another circumvention system, like Anonymizer.

Ease of deployment: We argue that SWEET can be easily deployed on the Internet and provide service to a wide range of users. First of all, SWEET is low-cost and needs few resources for deployment. It can be deployed using a single server that runs a few light-weight processes, including a mail server and a SOCKS proxy. To service in a large scale SWEET server can be deployed in a distributed manner as several machines in different geographic locations. Secondly, the operation of SWEET is standalone and does not rely on collaboration from other entities, e.g., end-hosts or ISPs. This provides a significant advantage to recent research that relies on collaboration from ISPs, or end-hosts. In fact, the easy setup and low-resources of SWEET's deployment allows it to be implemented by individuals with different levels of technical expertise. For instance, an ordinary home user can deploy a personal SWEET server to help her friends in censored regions evade censorship, or a corporate network can setup such system for its agents residing in a censored country.

User convenience: As a recent study shows, users give the most preference to the ease of use when choosing a circumvention tool. The use of SWEET is simple and requires few resources from a client. A SWEET client only needs to install the client software, which can be obtained from out-of-band channels like social networks or downloaded from the Internet. Due to its simple design, an expert user can also develop the client software herself. In addition, the user needs to have an email account with a public email provider, and needs to know the public information related to SWEET, e.g., the email addresses of SWEET.

VI. Prototype Implementation

A. Server Implementation

We implement the SWEET server on a Linux machine, which runs *Ubuntu 10.04 LTS* and has a 2 GHz quad-core CPU

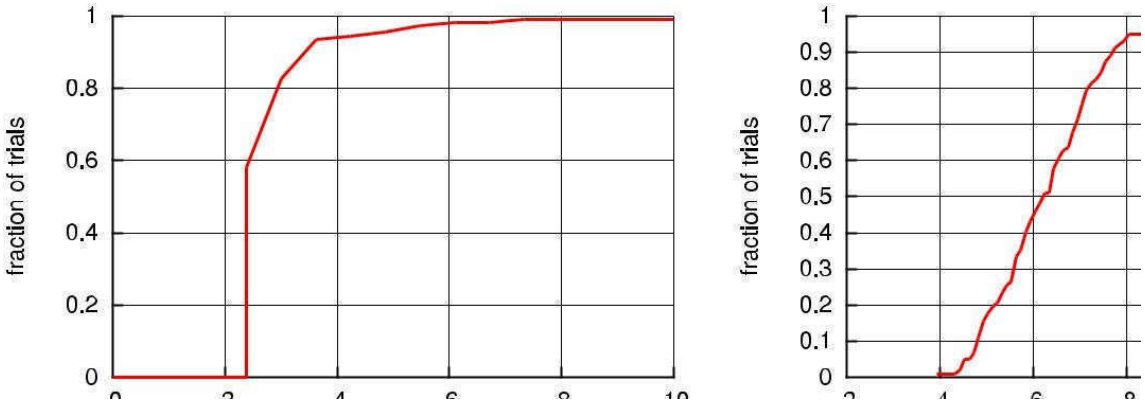


Fig. 5. The CDF of (a) the time that a SWEET email takes to travel from the SWEET client to the SWEET server; (b) the registration time.

and 4 GB of memory. We run Postfix,[1] a simple email server that supports basic functions. Postfix listens for new emails targeted to the `register@sweet.org` and `tunnel@sweet.org` email addresses. Postfix stores the received emails into designated file directories that are continuously watched by the converter and registration agent of SWEET server. Each stored email has a unique name, concatenating the email id of its corresponding client and an increasing counter. The converter agent is a simple Python-based program that runs in the background and continuously checks the folder for new emails. The converter also converts proxied packets, passed by SWEET's proxy, into emails and sends them to their intended clients. For the proxy agent, we use Squid[2] as our HTTP proxy and Suttree[3] as our SOCKS proxy. Squid listens on a local port for connections from the converter.

B. Client Implementation

Protocol-based design: The client prototype is built on a desktop machine, running *Linux Ubuntu 10.04 TLS*. We set up a web browser to use the local port "localhost:9034" as the SOCKS/HTTP proxy. The converter is a simple python script that listens on port 9034 for connections, e.g., from our web browser. We implement the email agent of SWEET client using *Fetchmail*,[4] a popular client software for sending and retrieval of emails through email protocols. We generate a free Gmail account and configure Fetch mail to receive emails through IMAP[5] and POP3[6] servers of Gmail, and to send emails through the SMTP server of Gmail.[7] Note that our design does not rely on Gmail, and the client software can be set up with any email account.

Webmail-based design: Our webmail-based implementation also runs on *Linux Ubuntu 10.04 TLS*, and uses the same converter as the one used in the protocol-based prototype.

A Google Chrome browser is used for making connections through SWEET, configured to use "localhost:9034" as a proxy. We prototype the web-based email agent by running a *UserScript*[8] inside the Mozilla Firefox[9] browser. More specifically, we install a Firefox extension, *Greasemonkey*,[10] to allow a user to run her own JavaScript, i.e., User script, while browsing certain destinations. We write a User Script that runs in Gmail's webmail interface and listens for the receipt of new emails. Our User Script saves new emails in a local directory, which is watched by the converter. Note that the Firefox browser is directly connected to the Internet and does not use any proxies (user needs to use the configured Chrome browser to surf the web through SWEET).

VII. Evaluation

We evaluate SWEET using our prototype implementation.

A. Performance

We use Gmail as the oblivious mail provider in our experiments. Our SWEET server is located in Urbana, IL, resulting in approximately 2000 miles of geographic distance between the SWEET server and Gmail's email server (we locate Gmail's location from its IP address). Fig. 5(a) shows the CDF of the time that a SWEET email (carrying the tunneled traffic) sent by a SWEET client takes to reach our SWEET server (the reverse path takes a similar time). As the figure shows, around 90% of emails take less than 3 seconds to reach the server, which is very promising considering the high data capacity of these emails. Note that based on our measurements, most of this delay comes from email handling (eg, spam checks, making SMTP connections, etc.) performed by the oblivious mail provider (Gmail in our experiments), but not from the network latency (the network latency and client latency constitute only tens of milliseconds of the total latency). As a result, the latency would be very similar for users with an even longer geographical distance from the oblivious mail server.

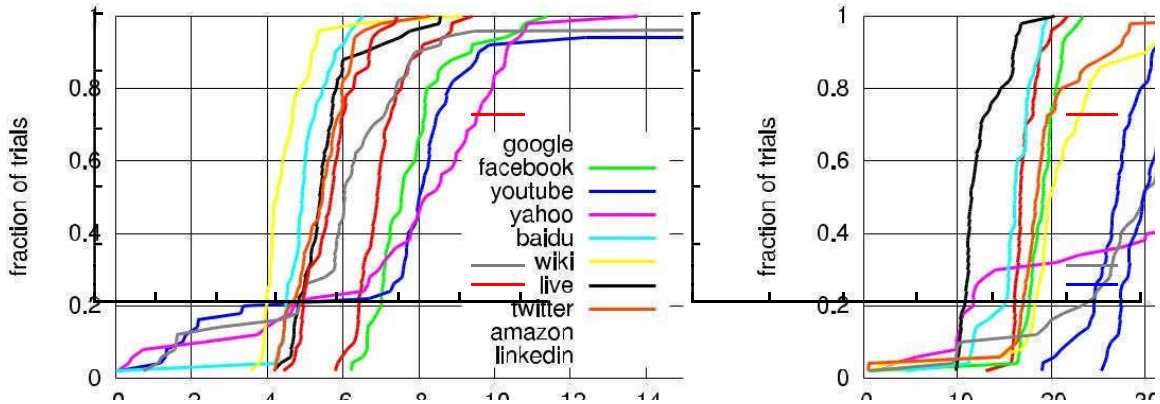


Fig. 6. The CDF of (a) the time to the first appearance (TFA) and (b) the total browsing time (TBT) using SWEET.

Client registration: Before being able to request data from Internet destinations, a user needs to be registered by the SWEET server. Fig. 5(b) shows the time taken to exchange registration messages between a client and the SWEET server. Note that the client registration needs to be performed *only once* for a long period of time. The figure shows that more than 90% of registrations establish in less than 8 seconds (with an average of 6.4 seconds).

We use two metrics to evaluate the latency performance of SWEET in browsing websites: the *time to the first appearance (TFA)* and the *total browsing time (TBT)*. The TFA is the time taken to receive the first response from a requested web destination. It is an important metric in measuring user convenience during web browsing. For instance, suppose that a client requests a URL, e.g., http://www.cnn.com/some_news.html. By the TFA time the client receives the first HTTP RESPONSE(s) from the destination, which include the URL's text parts (perhaps the news article) along with the URLs of other objects on that page, e.g., images, ads hosted by other websites, etc. At this time the client can start reading the received portion of the website (e.g., the news article), while her browser sends requests for other objects on that webpage. On the other hand, the total browsing time (TBT) is the time after which the browser finishes fetching all of the objects in the requested URL.

Using our prototype we measure the end-to-end web browsing latency for the client to reach different web destinations. Fig. 6(a) shows the TFA for the top 10 web URLs from Alexa's most-visited sites ranking [46]. The median is about 5 seconds across all experiments, which is very promising to user convenience.

On the other hand, Fig. 6(b) shows the total browsing time (TBT) for the same set of destinations (50 runs for each website). As can be seen, the destinations that contain more web objects (e.g., yahoo and linke din) take more time to get completely fetched (note that after the TFA time the user can start reading the webpage). We also run similar experiments through the popular Tor anonymous network to compare its latency performance with SWEET. Fig. 7 compares the latency CDF for SWEET and Tor. As expected, our simple implementation of SWEET takes more time than Tor to browse web pages, however, it provides a sufficient performance

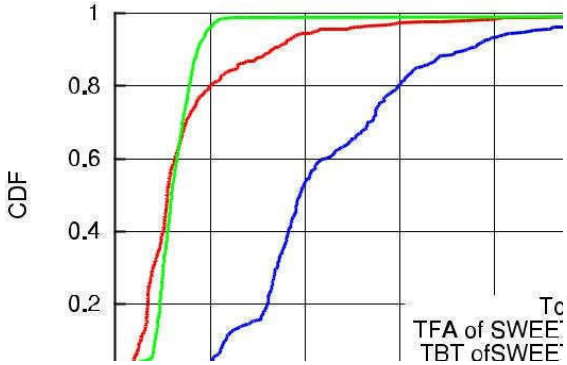


Fig. 7. Comparing the average latency of SWEET and Tor.

for normal web browsing. This is in particular significant considering the strong *availability* of SWEET compared to other circumvention systems. Additionally, we believe that further optimizations on SWEET server's proxy (like those implemented by Tor exit nodes) will further improve the performance. Our techniques are also amenable to standard methods to improve web latency, such as plugin-based caching and compression, which can make web browsing tolerable in high delay environments.

B. Traffic Analysis

A powerful censor can perform traffic analysis to detect the use of SWEET, e.g., by comparing a user's email communications with that of a typical email user. As a result, a SWEET user who is concerned about unobservability needs to ensure that her SWEET email communications mimic that of a normal user (a user who does not fear reprisal from her government might opt to have lower unobservability in order to gain a higher communication bandwidth). It should be mentioned that such traffic analysis is expensive for censors considering the large volume of email communications; it is estimated[11] that 294 billion emails were sent per day in 2011.

Fig. 8 shows the number of emails sent and received by a SWEET client to browse different websites. We observe that for any particular website the number of emails does not change at different runs. As can be seen, most of the web sites

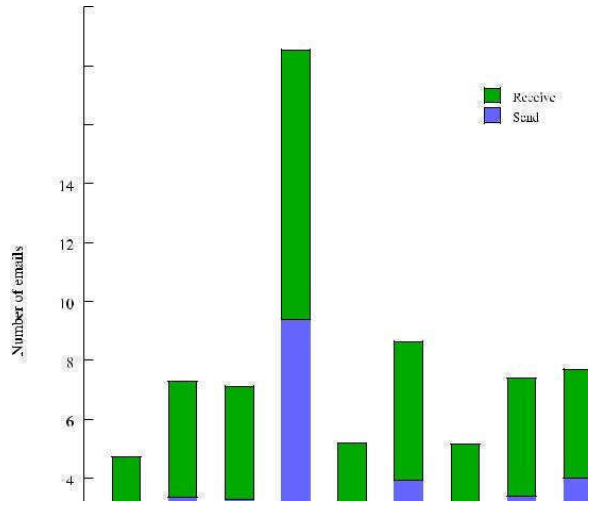


Fig. 8. The number of emails sent and received by a SWEET client to get one of the websites from Alexa's top ten ranking.

finish in less than three SWEET emails in each direction. The exception is the Yahoo web page as it contains many web objects, each hosted by different URLs (note that the number of emails affects the latency performance only *sub-linearly*, since some emails are sent and received simultaneously.). Also, the average number in each way of a connection is about 4 emails. A recent study on email statistics predicts that an average user will send 35 emails and will receive 75 emails per day in 2012 (the study predicts the numbers to increase annually). In addition, membership in mailing lists[12] and Internet groups[13, 14] is popular among Internet users, producing even more emails by normal email users. As an indication of the popularity of such services, Yahoo in 2010 announced[15] that 115 million unique users are collectively members of more than 10 million Yahoo Groups. Based on the mentioned statistics, we estimate that a conservative SWEET user can perform 35-70 web downloads per day, or make 10-20 interactive web connections, while ensuring unobservability of SWEET usage. Note that the censored users use SWEET only to browse "censored" Internet webpages and they use regular web browsing for non-censored websites. Also, normal citizens who do not fear being caught by the censors may decide to ignore resistance against traffic analysis in order to achieve higher bandwidths.

The censors may also try to detect SWEET users by analyzing the inter-arrival times of the email messages exchanged between SWEET users and their mail service providers. More specifically, a SWEET client may send and receive multiple emails in a shorter time interval compared to regular email clients. We, however, argue that this is not a serious vulnerability for SWEET. As discussed earlier, we require SWEET clients to use mail service providers that encrypt email exchanges (otherwise, the censors can simply filter SWEET emails by searching for the email addresses of SWEET servers). The use of encryption prevents the censors from identifying the number of emails exchanged between a SWEET user and her mail service provider.

VIII. Conclusions

In this paper, we presented SWEET, a deployable system for unobservable communication with Internet destinations. SWEET works by tunneling network traffic through widely-used public email services such as Gmail, Yahoo Mail, and Hotmail. Unlike recently-proposed schemes that require a col-lection of ISPs to instrument router-level modifications in sup-port of covert communications, our approach can be deployed through a small applet running at the user's end host, and a remote email-based proxy, simplifying deployment. Through an implementation and evaluation in a wide-area deployment, we find that while SWEET incurs some additional latency in communications, these overheads are low enough to be used for interactive accesses to web services. We feel our work may serve to accelerate deployment of censorship-resistant services in the wide area, guaranteeing high availability.

References

[1] J. Zittrain and B. Edelman, "Internet filtering in China," *IEEE Internetcc Comput.*, vol. 7, no. 2, pp. 70–77, Mar. 2003.
[2] (Nov. 2007). *Defeat Internet Censorship: Overview of Advanced Technologies and Products*. [Online]. Available: http:// www.internetfreedom.org/archive/DefeatInternet Censorship White Paper.pd
[3] C. S. Leberknight, M. Chiang, H. V. Poor, and F. Wong. (2010). *A Taxonomy of Internet Censorship and Anti-Censorship*. [Online]. Available: http://www.princeton.edu/ chiangm/anticensorship.pdf
[4] I. Clarke, S. G. Miller, T. W. Hong, O. Sandberg, and B. Wiley, "Protecting free expression online with freenet," *IEEE Internet Comput.*, vol. 6, no. 1, pp. 40–49, Jan. 2002.
[5] *Ultrasurf*, accessed on Jan. 7, 2017. [Online]. Available: https://ultrasurf.us/
[6] J. Jia and P. Smith. (2004). *Psiphon: Analysis and Estimation*. [Online]. Available: http://www.cdf.toronto.edu/ csc494h/reports/2004-fall/psiphon_ae.html
[7] I. Cooper and J. Dilley, "Known HTTP proxy/caching problems," IETF, Fremont, CA, USA, Tech. Rep. Internet RFC 3143, Jun. 2001.
[8] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," in *Proc. USENIX Secur. Symp.*, 2004, pp. 21–37.
[9] J. Boyan, "The anonymizer: Protecting user privacy on the Web," *Comput.-Mediated Commun. Mag.*, vol. 4, no. 9, pp. 1–6, Sep. 1997.
[10] *DynaWeb*, accessed on Jan. 7, 2017. [Online]. Available: http://www.dongtaiwang.com/home_en.php
[11] R. Clayton, S. J. Murdoch, and R. N. M. Watson, "Ignoring the great firewall of China," in *Proc. Int. Workshop Privacy Enhancing Technol.*, 2006, pp. 20–35.
[12] Y. Sovran, A. Libonati, and J. Li, "Pass it on: Social networks stymie censors," in *Proc. 7th Int. Conf. Peer-to-Peer Syst.*, Feb. 2008, p. 3. [Online]. Available: http://www.iptps.org/papers-2008/73.pdf
[13] D. McCoy, J. A. Morales, and K. Levchenko, "Proximax: A measure-ment based system for proxies dissemination," *Financial Cryptogr. Data Secur.*, vol. 5, no. 9, pp. 1–10, 2011.
[14] N. Feamster, M. Balazinska, W. Wang, H. Balakrishnan, and D. Karger, "Thwarting Web censorship with untrusted messenger discovery," in *Int. Workshop Privacy Enhancing Technol.*, 2003, pp. 125–140.
[15] M. Mahdian, "Fighting censorship with algorithms," in *Proc. Int. Conf. Fun Algorithms*, 2010, pp. 296–306. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13122-6_29
[16] J. McLachlan and N. Hopper, "On the risks of serving whenever you surf: Vulnerabilities in Tor's blocking resistance design," in *Proc. 8th ACM Workshop Privacy Electron. Soc.*, Nov. 2009, pp. 31–40. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1655188.1655193
[17] P. Winter and S. Lindskog. Apr. 2012. "How China is blocking Tor." [Online]. Available: https://arxiv.org/abs/1204.0447
[18] (Sep. 2007). *Tor Partially Blocked in China*. [Online]. Available: https://blog.torproject.org/blog/tor-partially-blocked-china

[19] N. Feamster, M. Balazinska, G. Harfst, H. Balakrishnan, and D. Karger, "Infranet: Circumventing Web censorship and surveil-lance," in *Proc. 11th USENIX Secur. Symp.*, Aug. 2002, pp. 247–262. [Online]. Available: http://www.usenix.org/events/sec02/feamster.html

[20] S. Burnett, N. Feamster, and S. Vempala, "Chipping away at censorship firewalls with user-generated content," in *Proc. USENIX Secur. Symp.*, 2010, pp. 463–468. [Online]. Available: http://www.usenix.org/events/sec10/tech/full_papers/Burnett.pdf

[21] R. Dingledine and N. Mathewson, "Design of a blocking-resistant anonymity system," Tor Project, Tech. Rep. 1, Nov. 2006.

[22] J. Karlin *et al.*, "Decoy routing : Toward unblockable Internet commu-nication," in *Proc. USENIX (FOCI)*, 2011, pp. 1–6.

[23] E. Wustrow, S. Wolchok, I. Goldberg, and J. A. Halderman, "Telex: Anticensorship in the network infrastructure," in *Proc. 20th USENIX Secur. Symp.*, Aug. 2011, pp. 1–15.

[24] A. Houmansadr, G. T. K. Nguyen, M. Caesar, and N. Borisov, "Cirri-pede: Circumvention infrastructure using router redirection with plau-sible deniability categories and subject descriptors," in *ACM Conf. Comput. Commun. Secur. (CCS)*, Chicago, IL, USA, 2011, pp. 187–200.

[25] H.-C. Hsiao *et al.*, "LAP: Lightweight anonymity and privacy," in *Proc. IEEE Symp. Secur. Privacy*, May 2012, pp. 506–520.

[26] H. M. Moghaddam, B. Li, M. Derakhshani, and I. Goldberg, "Skype-morph: Protocol obfuscation for Tor bridges," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, 2012, pp. 97–108.

[27] Q. Wang, X. Gong, G. Nguyen, A. Houmansadr, and N. Borisov, "Censor Spoofer: Asymmetric communication using IP spoof-ing for censorship-resistant Web browsing," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, 2012, pp. 121–132.

[28] Z. Weinberg *et al.*, "Stego Torus: A camouflage proxy for the Tor anonymity system," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, 2012, pp. 109–120.

[29] A. Houmansadr, C. Brubaker, and V. Shmatikov, "The parrot is dead: Observing unobservable network communications," in *Proc. IEEE Symp. Secur. Privacy*, May 2013, pp. 65–79.

[30] A. Houmansadr, T. Riedl, N. Borisov, and A. Singer, "I want my voice to be heard: IP over voice-over-IP for unobservable censorship circumvention," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2013, pp. 1–17.

[31] M. Leech *et al.* (Apr. 1996). *RFC 1928: SOCKS Protocol Version 5*. [Online]. Available: http://ftp://ftp.internic.net/rfc/ rfc1928.txt,ftp://ftp.math.utah.edu/pub/rfc/rfc1928.txt

[32] C. Brubaker, A. Houmansadr, and V. Shmatikov, "Cloud transport: Using cloud storage for censorship-resistant networking," in *Proc. PETS*, 2014, pp. 1–20.

[33] H. Zolfaghari and A. Houmansadr, "Practical censorship evasion lever-aging content delivery networks," in *Proc. 23rd ACM Conf. Comput. Commun. Secur. (CCS)*, 2016, pp. 1–12.

[34] J. Holowczak and A. Houmansadr, "CacheBrowser: Bypassing Chinese censorship without proxies using cached content," in *Proc. 22nd ACM Conf. Comput. Commun. Secur. (CCS)*, 2015, pp. 70–83.

[35] R. McPherson, A. Houmansadr, and V. Shmatikov, "CovertCast: Using live streaming to evade Internet censorship," in *Proc. Privacy Enhancing Technol. (PETS)*, 2016, pp. 212–225.

[36] M. Schuchard, J. Geddes, C. Thompson, and N. Hopper, "Routing around decoys," in *Proc. ACM Conf. Comput. Commun. Secur.*, 2012, pp. 85–96.

[37] A. Houmansadr, E. L. Wong, and V. Shmatikov, "No direction home: The true cost of routing around decoys," in *Proc. 21st Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2014, pp. 1–14.

[38] M. Nasr and A. Houmansadr, "GAME OF DECOYS: Optimal decoy routing through game theory," in *Proc. 23rd ACM Conf. Comput. Commun. Secur. (CCS)*, 2016, pp. 1–12.

[39] *The FOE Project*, accessed on Jul. 1, 2013. [Online]. Available: http://code.google.com/p/foe-project/

[40] *MailMyWeb*, accessed on Jan. 7, 2017. [Online]. Available: http://www.mailmyweb.com/

[41] A. Juels and J. G. Brainard, "Client Puzzles: A cryptographic counter-measure against connection depletion attacks," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 1999, pp. 151–165. [Online]. Available: http://www. isoc.org/isoc/conferences/ndss/99/proceedings/papers/juels.pdf

[42] M. Steiner, G. Tsudik, and M. Waidner, "Diffie–Hellman key distribution extended to groups," in *Proc. 3rd ACM Conf. Comput. Commun. Secur.*, Mar. 1996, pp. 31–37.

[43] R. Fielding *et al.*, "Hypertext transfer protocol—HTTP/1.1," IETF, Fre-mont, CA, USA, Tech. Rep. RFC 2616, Jun. 1999, [Online]. Available: http://www.ietf.org/rfc/rfc2616.txt

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Student Intervention System using Machine Learning Techniques

*

Shubhangi Urkude[1]
[1]*Department of Computer Science and Engineering*
Faculty of Science and Technology
The ICFAI Foundation for Higher Education (IFHE)
Hyderabad, Telangana, India – 501203.
ushubhu@ifheindia.org

Kshitij Gupta[2]
[2]*Department of Computer Science and Engineering*
Faculty of Science and Technology
The ICFAI Foundation for Higher Education (IFHE)
Hyderabad, Telangana, India – 501203.
cool.viju1721@gmail.com

Abstract*— Now a days, the educational institutes are adopting technologies for betterment of student's quality, in respect to teaching methodologies etc. For which the huge information available with educational institutes can be used to predict student's future in academics. The main objective of this paper is to predict the student performance in the examination and also to predict the student will graduate or not. Hence forth we are using statistical analytical method which is F1 score. F1 score or F measure is used to test the prediction accuracy by considering precision and recall to compute the score. To fulfill this requirement in machine learning, classification technique is used. The dataset used in this analysis contains 395 student records, having attributes, such as age, health, internet, school, father job, mother job etc. Using support vector machines (SVM), Decision Tree and Naïve Bayes (NB) classification algorithms F1 score is calculated for each algorithm. Based on the analysis done the F1 score of support vector machine is giving the better prediction compared to rest of the two algorithms.

Keywords*— Support Vector Machines (SVM), Decision Tree, Naïve Bayes (NB), Classification Algorithm, Prediction.

I. INTRODUCTION

In educational institutes like colleges and many universities huge data is collected on daily basis. This data contains examination data, student's internal marks, their personal data, health data, various evaluation components data and logs of student's activity. The data is caught through learning administration frameworks like Canvas and Edmodo. Canvas and Edmodo are particularly intended for online classrooms, which are getting to be well known in the higher education. This huge collection of data can be used to analyze the student's performance in the future to increase the graduation rate. Graduation rates are frequently the criteria of decision for this. Graduation rate is defined as the time when student enters the institution and completed the degree with in four years, it is usually expressed as % graduation rate. To finalize the graduation rate the data collected from different sources should be converted into single form so that this data can be retrieved and analyzed to make specific decision. To take any decision, data is analyzed in such a way that it should able to answer certain questions such as

Who are the students likely to fail?
What is the courses student may fail?
What is the quality of the student's contribution to the marks he obtained?
Which courses should be focused more to help the student?
What are the easy subjects to get good score?

Based on the answers collected in the analysis phase, the perfect decision can be made by using different machine learning techniques and students can be guided in a particular direction without diluting educational standards.

Student's academic performance influenced by numerous components, like personal, financial and other environmental variables. Learning about these factors and their impact on student performance can help dealing with their impact. Recently, much aid has been paid to educational mining research. Educational Data Mining (EDM) concerns techniques, tools, and research planned for automatically extracting meaning from big repositories of data generated by or related to people learning activities in educational surrounding. Anticipating student's performance turns out to be all the more challenging because of the expansive volume of information in educational databases.

Forecasting of academic performance is widely explored. The ability to foresee student performance is important in educational institutions. Expanding student's achievement is a long-haul objective in every single educational institute. Predicting student's academic performance ahead of schedule before their last examination is really a great achievement. At this point additional measures can be taken by educational institutes to orchestrate appropriate help for the low performing students to enhance their studies and help them to progress. On the other hand, distinguishing characteristics that influence course achievement rate can aid courses enhancement. Recently created web-based educational technologies and the use of quality standard offer researchers' novel chances to contemplate how students realize and what ways to deal with learning lead to progress. The main objective of this paper is to distinguish the two factors that influence student's performance, courses achievement rate and student's achievement rate. By considering these components as early indicator analysis can be done and necessary action can be taken to improve.

II. LITERAYURE SURVEY

Many researchers have used statistics and machine learning algorithms for predicting the student's performances in educational institutes.

Edin Osmanbegovic et al. [1] uses three supervised learning algorithms, Bayesian, Decision trees and Neural Networks on the assessment data to predict number of students got success in a course and the performance of the teaching, learning methods were evaluated based on their predictive precision and ease of learning and user-friendly features. It is proved that this method can be used to benefit

students and teachers to progress student's performance and decrease failing ratio by taking proper steps at right time to improve the quality of education.

Mladen Dragicevic, et al. [2] used decision tree classification to predict student's performance based on Grade Point Average (GPA) criteria. Two parameters are considered for analysis as time-in-degree and GPA, in that GPA is giving better result in predicting the student's performance.

CH.M.H.Sai Baba, et al. [3]used decision tree and multivariant regression analysis to predict the number of students getting job.

Behrouz MinaeiBidgoli I et al. [4] compared four different classifiers and united the results into a multiple classifier. Their research separated the data into three diverse classes as high, middle and low. Genetic algorithm is used to improve the accuracy of prediction in all the classes. For the data set having less features, feature weighting mechanism is better as compare to feature selection. Using LON-CAPA, e-learning platform results are validated.

Furthermore, In the research work Rahel Bekele et al. [5], applied Bayesian network in the education field, to forecast student's performance. This model is also tested on real world data where students are assigned t fill the data and analysis is done on the actual data to predict their performance. It shows that results obtained are very useful for the teacher to assist the student to improve the academic performance.

Paulo Cortez et al. [6] addressed the prediction of secondary school student's performance in two core subjects as mathematics and Portuguese by using their past score in the earlier session and other demographic factors. Using Business Intelligence (BI) and Data mining techniques such as Decision trees, Random Forests, Support Vector machines method and Neural network real world data is analyzed. The real data may contain information related to student's grade, social and school related features. The model is tested with and without previous semester grades. Model is showing good predictive accuracy

Surjeet Kumar Yadav et al. [7], concluded that Decision trees are so popular because they produce classification rules that are easy to interpret than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for Student data to predict the student's performance in the end semester examination. The experimental results show that Classification and Regression Tree (CART). CART is the best algorithm for classification of data. It is giving the accuracy of 56.25% as compare to other algorithms like ID3 and C4.5. This study is useful for the students who are going to fail, and also used to identify the poor student which requires more attention.

From the study conducted by Zlatko J. Kovacic [8], proposed a case study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success. Two algorithms CHAID and CART were applied on student enrolment data of information system students of open polytechnic of New Zealand to get two decision trees classifying, successful and unsuccessful students. The result obtained showed that the accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

AbeerBadr El Din Ahmed et al. [9] used classification technique to predict the final grade of students. It was done by the use of ID3 decision tree method. To determine best attribute for particular node in the tree Information Gain measure is used over the collected sample S. Midtrem attribute is having highest gain, so it is selected as root node for decision tree and the same process is followed to classify entire attributes.

Sembiring S et al. [10] showed that Data Mining Techniques (DMT) capabilities provided effective improving tools for student performance. The study further showed how useful data mining can be in higher education particularly to predict the final performance of student. The researchers gathered data from student by using questionnaire to find the relationships between behavioral attitude of student and their academic performance. Data mining techniques were then applied. They obtained the prediction rule model using decision tree as well as applying the rules into Support Vector Machine (SVM) algorithm to predict the student's final grade. Also, the students were clustered into groups using kernel k-means clustering. The study expressed the strong correlation between mental condition of student and their final academic performance.

Ogunde A. O, et al. [11] used Iterative Dichotomiser 3 (ID3) decision tree algorithm to predict the university student's grade in Nigeria. A prediction accuracy of 79.556% was obtained from the model. They recommended different decision-tree model to perform similar analysis with extended data set to get better results.

Dorina Kabakchieva et al. [12] proposed classification method to predict student performance. This paper compares different data mining algorithms using WEKA tool and results are likely to vary between 52%-67%. For the analysis university sample data related to student is collected containing admission score, number of failures at the first year etc.

HashmiaHamsa, et al. [13] proposed academic performance model using decision tree and fuzzy genetic algorithm. In this research work different parameters like internal marks, attendance, average marks etc. are considered to identify the student's performance in degree and master's degree students. Degree student's performance is evaluated by using decision tree algorithm which gave more students at risk. Fuzzy genetic algorithm giving more pass students by considering students in between risk and safe state.

III. MOTIVATION

However, with constrained assets and budgets, the leading body of administrators needs us to locate the best model with minimal measure of calculation costs. With the end goal to build the intervention software, we first should break down the dataset on student's execution. Our goal is to choose and build up a model that will foresee the probability that a given student will pass, along diagnose whether or not an intervention is necessary. Our model is developed based on a subset of the information that we give, and it will be tested against a subset of the information that is kept hidden from the learning algorithm, with the end goal to test the model's effectiveness on information outside the training set.

ISBN 978-81-939386-1-4

IV. PROBLEM STATEMENT

The college administration a has an objective to achieve a 95% graduation rate before the decade's over by distinguishing students who require intervention before they drop out of school. To recognize whether the student will graduate or not. We will probably display the variables that foresee how likely a student is to pass their secondary school final exam. We being a sharp engineer choose to implement a student intervention framework utilizing ideas we gained from supervised machine learning. Rather than purchasing costly servers or actualizing new information models starting from the earliest stage, it can be connected to outsider organization who will give the essential programming libraries and servers to run our software.

V. MACHINE LEARNING ALGORITHM

Unsupervised learning

At the point when the dataset isn't characterized or hard for interpretation, it is called unsupervised learning. The labels for the information are not characterized. There no correct method to partition information collected except performing iterations. As per this research problem, unsupervised learning isn't advisable for prediction.

Supervised learning

Supervised learning can be said as function approximation, in which training examples lead to function generation. If the learning is finished with right training set, a well-behaved function can be expected. Supervised learning develops reliably with the data. It is a sort of induction learning, and it causes one-sided supervised learning.

E.g.: The function generated with supervised learning will be $X2$, if X is simply the information esteem and the output is self-duplicated.

Classification algorithm

In machine learning and measurements, classification is the issue of distinguishing in different labels, based on a training set of information containing observations (or occurrences) whose category is known. Classification is technique where details arranged into a given number of classes. Classifier: A algorithm that maps the input data to a particular classification. A classification model tries to make some inference from the input values given for training.

VI. DATA COLLECTION AND PREPARATION PHASE

The dataset was collected from the website Kaggle.com. This dataset consists of 395 student records; each record consists of 30 attributes with their domain values. The attributes are listed in Table1. The dataset was divided two parts, training dataset (75%) and testing dataset (25%).

Attribute Name	Description
School	General public
Sex	Male or female
Age	Between 15-18
Address	Urban or rural
Family size	Numeric value (1-5)
Parental status	Yes or no
Mother Education	Numeric value (1-5)
Father education	Numeric value (1-5)
Mother job	At_home, services, other
Father job	Teacher , services and other
Reason	Course or home
Gaurdian	Mother or father
Travel time	Numeric value (1-5)
Study time	Numeric value (1-5)
Faliures	Numeric value (1-5)
Paid	Yes or no
Activities	Yes or no
Nursery	Yes or no
Higher	Yes or no
Internet	Yes or no
Romantic	Yes or no
freetime	Numeric value (1-5)
Go out	Numeric value (1-5)
Health	Numeric value (1-5)
Absences	Numeric value (1-5)
Passed	Yes or no
Schools up	Yes or no
Families up	Yes or no
femeral	Numeric value (1-5)
Dalc	Numeric value (1-5)
Walc	Numeric value (1-5)

Table 1: List of attributes

VII. CLASSIFICATION MODELS

The classification algorithms used in this research is

Support vector Machine (SVM)

SVM are supervised learning models that analyze data and distinguish patterns and used for classification analysis. Some of its advantages are SVM is very efficient in high-dimensional spaces, and in situations when we have a non-linear separation problem. With SVM we have the probability to apply new kernels that allows tractability for our decision boundaries, contributing to a better classification performance. One major disadvantage of the SVM is the deciding the kernel and also it is slow when compared to other models like Decision trees or Naive Bayes.

Decision Trees

The advantages of a decision trees are that nonlinear relationships between parameters do not influence our performance metrics and they give us quicker prediction as compared to other models like SVMs. Decision Trees do not function well if we have smooth boundaries. i.e they work best when we have discontinuous piece wise constant model. If we really have a linear target function decision trees are not the best.

Naive Bayes

Naive Bayes is one of the quickest learning algorithms. It is mainly used for classic problems such as spam detection, recognizing letters from handwritten texts and facial analysis. The advantages of naive Bayes is that it is faster than other discriminative models like logistic regression, so we need less training data. The same conditional independence assumption can be a drawback when we have no occurrences of a class label and a feature value untidily, that will give us a zero frequency-based value probability that affects any posterior probability estimate.

These are the Classification models that were assessed and all had promising outcomes. Analyzed outcomes can be utilized in enhancing student's academic performance. It is pointed out that the corrective action must be the joint endeavors of educators, guardians/parents, the advanced education organization and the students. Some corrective measures that are distinguished incorporate support and inspirations, nearly checking the students and extra or extramural instructional exercises.

VIII. RESULT

This segment will examine the outcomes analysis of the ongoing works in predicting student's performance. This Meta investigation depends on the highest accuracy of prediction methods and furthermore the primary vital components that may impact the student's performance. F1 score measure is taken to find student's performance. F1 score achieves its best value at 1 and worst at 0. For finding the best model among the selected models first sample data of 100 students was tested and F1 score is calculated, then sample size 200 students were taken and finally same model is tested with sample size of 300 records. For each sample size F1 score is calculated. Figure 1 shows the F1 score for decision tree. In this for a sample size of 100 records F1 score is 0.6552, for 200 records F1 score is increase to 0.75 and for sample of 300 records it is decreases to 0.6613. From this plot it is understood that performance of decision tree is increase up to certain input limit and then it suddenly decreases to initial score.

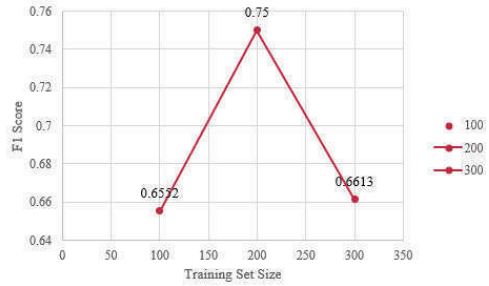


Figure 1: F1 Score for Decision Tree

Figure 2 shows the F1 score for Naïve Bays technique. In Naïve Bays a sample size of 100 records giving F1 score as 0.8029, for 200 records F1 score is decrease to 0.7244 and for sample of 300 records it is increase to 0.7634. From this plot it is understood that performance of decision tree is excellent for a smaller number of inputs.

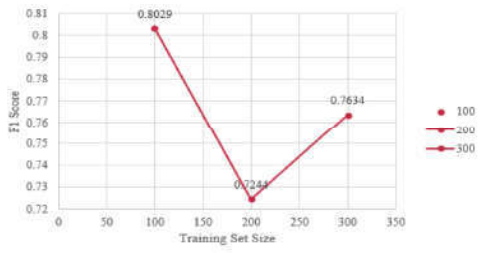


Figure 2: F1 Score of Naïve Bayes

Figure 3 shows the F1 score for Support Vector Machine technique. In this a sample size of 100 records giving F1 score as 0.7746, for 200 records F1 score is increase to 0.7815 and for sample of 300 records it is increase to 0.7638. From this plot it is understood that performance of support vector machine is increasing with data set size. So, for analyzing the student's performance this is best model as number of student's intake is increases prediction accuracy also increase.

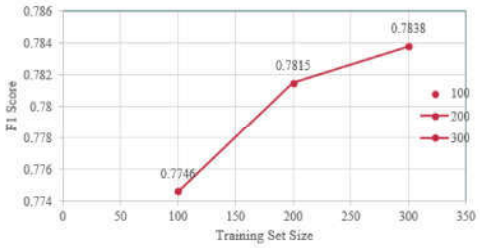


Figure 3: F1 Score for Support Vector Machine

Finally, all machine learning algorithms are compared with same input records and comparison table is prepared. Table 2 shows the F1 score for all algorithms with their accuracy for training data set and testing data set.

Algorithms	Decision Tree	Naïve Bayes	Support Vector Machine
Training set size	300	300	300
F1 score (Training set)	1.000	0.8038	0.8761
F1 score (Testing set)	0.6613	0.7634	0.7838

Table 2: Final F1 score for all algorithms

Figure 4 shows the graphical representation for all algorithms with F1 score value. According to the analysis support vector machine is giving the best result.

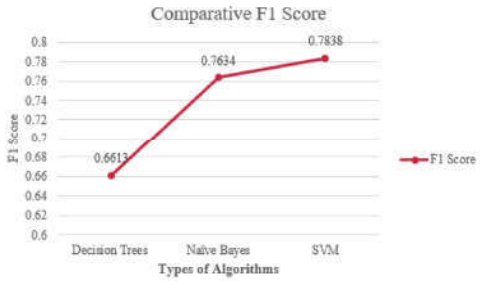


Figure 4: Final F1 Score Comparison for student intervention system

IX. CONCLUSION AND FUTURE SCOPE

This study of research paper thought in the beginning endeavor to utilize machine learning techniques to break down and discover student academic performance and to enhance the nature of the education framework. The administrations can utilize these systems to enhance the course results to enhance the student performance. Such information can be utilized to give a decent comprehension of student's enrollment pattern in the courses under study. The faculty and managerial decision maker can utilize this research to improve performance. Other hand, such sort of learning encourages the administration to upgrade their policies, enhance their strategies and enhance the quality of the system.

In this research Support vector machines, Decision Trees, Naïve Bayes algorithm techniques are utilized to predict student's learning activities. We trust that the information created might be useful for educator and also for students. This investigation enhances student's performance; reduce

failing ratio by making suitable strides at correct time to enhance the quality of education. In Future this kind of analysis can be implemented with a bigger dataset and more background factors of student at the first-year academic level. This study concentrated on deciding to what degree family background factors and past scholastic accomplishments influence student's first year academic performance, recognizing the particular machine learning algorithm that best model to predict student academic performance.

X. REFERENCES

[1] Osmanbegovic E. and Suljic M., "Data mining approach for predicting student performance",Journal of Economics and Business,Vol. X, Issue 1, 2012.

[2] MladenDragicevic, Mirjana Pejic Bach, and VanjaSimicevic, "Improving University Operations with Data Mining: Predicting Student Performance", International Journal of Social, Behavioral, Educational, Economic and Management Engineering Vol. 8, Issue 4, 2014.

[3] CH.M.H.Sai Baba, AkhilaGovindu,Mani Krishna Sai Raavi, and VenkataPraneethSomisetty, "Student Performance Analysis Using Classification Techniques", International Journal of Pure and Applied Mathematics, Vol. 115, No. 6, pp. 1-7, 2017.

[4] Behrouz M, Karshy D, Korlemeyer G and Punch W., "Predicting student performance: an application of data Mining methods with the educational web-based system", IEEE Frontiers in Education Conference,2003.

[5] BekeleR. andMenzelW. "A bayesian approach to predict performance of a student (BAPPS): A Case with Ethiopian Students", Journal of Information Science 2016.

[6] Cortez P and Silva A., "Using data mining to predict Secondary school student performance", Journal of information science, Vol. 2, issue 6, 2013.

[7] Surjeet K, Yadav, Bharadwaj B and Pal B., "Data Mining Applications: A comparative Study for Predicting Student's performance", International journal of innovative technology & creative engineering, Vol. 1, issue 12,2012.

[8] Kovacic Z., "Early prediction of student success: Mining student enrollment data", Informing Science & IT Education Conference,pp. 647-665, 2010.

[9] Ahmed A. B. E and Ibrahim S. E.,"Data Mining: A prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology, Vol. 2, issue 2, pp. 43-47,2014.

[10] Sembiring S, Zarlis M, Hartama D., Ramliana S andElvi W., "Prediction of student academic performance by an application of data mining techniques.", International Conference on Management and Artificial Intelligence, Vol.6, pp. 110-114,2011.

[11] Ogunde A.O. andAjibade D.A., "A data Mining System for Predicting University Students Graduation Grade Using ID3 Decision Tree approach", Journal of Computer Science and Information Technology, Vol. 2, No 1, pp. 01-26,2014.

[12] DorinaKabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification", Cybernetics and Information Technologies, Vol. 13, No 1, pp. 61-72, 2013.

[13] HashmiaHamsa, SimiIndiradevi and Jubilant J. K., "Academic Performance Model Using Decision Tree and Fuzzy Genetic Algorithm",International Conference on Recent Advancement and Effectual Researches in Engineering, Science and Technology, pp. 326-332, 2016.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Robust optimized DWT-SVD based Watermarking

Swaraja.K[1]
PG Scholar, CSE,
GRIET, Hyderabad,
Telangana

Meenakshi.K[2]
Professor, ECE
GRIET, Hyderabad,
Telangana

Padmavathi.Kora[3]
Professor, ECE
GRIET, Hyderabad,
Telangana

G.Karuna[4]
Professor, CSE,
GRIET, Hyderabad,
Telangana

Abstract— A video watermarking technique is projected in this work by exploiting discrete wavelet transform (DWT) in addition to Singular Value Decomposition (SVD) transform. In this paper, DWT is applied on every luminance frame which is divided into 8x8 blocks of the video 'V' thus producing dissimilar frequency sub-bands, later SVD transform is implemented on the selected dwt blocks of HH bands of all frames. The selected dwt blocks were acquired through exploiting the Artificial Bee colony Optimization algorithm (ABC). At the receiving part, retrieval of watermark contents is achieved by a similar evaluation scheme practiced during the embedding procedure. This scheme recommends better visibility of watermarked video along with the retrieval of watermark content due to DWT-SVD- ABC combination. Simulation results show that the imperceptibility of watermarked video is better as PSNR is above 53 dB for all set of videos. Besides, robustness of the scheme is better than the existing scheme for similar set of videos in terms of normalized correlation (NC).

Keywords—SVD transform, DWT, ABC, Imperceptibility, Robustness.

1.INTRODUCTION

With the progress of transaction of data on the Internet and the widespread practice of digital media which is economical, there have been of enormous curiosity to digital data owners in multimedia safety and multimedia copyright security. The digital commodity Proprietors are worried regarding illicit doubling of their artifacts. Safety and copyright fortification are becoming imperative issues in multimedia applications. Watermarking digital image/video is a procedure utilized for copyright security of digital media. Watermark is a data related to owner concealed into the digital media without disturbing its perceptual quality. In the event of any disagreement, the watermark content can be identified or extorted, moreover utilized as an evidence of ownership. The techniques in watermarking that do not entail the original image for extraction of the watermark are called "blind" watermarking. The main efficacies of the digital watermarking technology can be characterized into four broad categories [1] such as Copyright fortification, continuous checking, Validation, Security and Indiscernible Communications. Digital watermarking persists to survive even after acquiring the data at the receiver side. If subtle data is disclosed to illicit personal, the digital watermark enclosed in them can be utilized to trace back to the intended receiver [2]. Digital watermarking exercised for concealed communication augments an additional level of protection contrast to cryptography. Though, the assailant is conscious of the endurance of the secret key, can be confident that within adequate interval he can decode the information, whereas while inserting the watermark digitally, the assailant can in no way be assured that hidden content is being conveyed. The foremost benefit of digital watermarks is that it lasts to persist though the receiver acquires the material. In applications related to medical [3] , watermarks are accomplished for credentials and retrieval of distinct patient archives. This specific application may avert human faults such as record discrepancy consequently precluding mortal faults [2].

The watermark in the projected method is entrenched in the video clip frames. Methodologies in video watermarking can be ordered into two leading classes based on the process of concealing watermark data in the original video. The two classes are: Spatial domain watermarking as well as transform-domain watermarking [4-16]. In spatial-watermarking procedures, entrenching and recognition are accomplished based on values of spatial pixels of the entire video frame. Spatial-domain schemes are simple to execute; on the contrary they are not resilient in conflict to accustomed digital signal processing functions such as video coding. Watermarking in the transform domain is more protected and strong. Transform-domain schemes, in contrast, amend spatial pixel values of the host video in relation to encoded transform. Recurrently exercised transforms are the Discrete Cosine Transform (DCT) [7, 8], the Fast Fourier Transform (FFT), the Discrete Wavelet Transform (DWT) and the Singular Value Decomposition (SVD) [9-11]. Transform-domain schemes of watermarking are established to be more strong besides unnoticeable when assessed in contradiction to spatial domain methods, subsequently dispersing the watermark in the spatial domain of video frame, makes it very tough to eliminate the inserted watermark. Every transform has its specific characteristics and indicates the image in diverse ways. SVD and Schur decompositions [12] are two mathematical tools used to analyze matrices. When some perturbation occurs in the watermarked image, the extraction of the watermark is not affected much. SVD is computationally expensive. Watermarking by means of Schur decomposition is faster when weighed against to SVD decomposition. With the aim of developing the achievement of image processing schemes at present many nature

influenced optimization methods [13-15] have appeared as powerful tools. In the proposed work also we are exploiting the ABC method which is optimization scheme to obtain improved results concerned to Imperceptibility and robustness.

Herein a resilient unsighted digital video watermarking algorithm is projected. The anticipated scheme merged the SVD transform by the DWT transform besides exploiting the ABC method to yield a competent technique with more robustness besides imperceptibility. This paper is systematized as follows. Different methodologies such as DWT and SVD transform along with ABC algorithm is deliberated in section 2. Projected algorithm is presented in section 3. Simulation outcomes are presented and compared with the existing methods in section 4. Conclusions are specified in section 5.

2. METHODOLOGIES

The disparate theories and practices exploited in this work are deliberated in this section so as to acquaint the reader with the elementary concepts.

2.1. Singular value decomposition (SVD)

A symmetric matrix is fragmented by SVD into three sub-matrices in which singular values are alienated into diagonal matrix [16]. This progression is also recognized as diagonalization of matrix. The three matrices that have undergone decomposition are termed as left singular matrix (L), singular matrix (W) and right singular matrix (R). Assuming C as a symmetric matrix SVD can be calculated with Eq. 1.

$$C = LWR^T \quad (1)$$

In Eq (1), Matrix W appear as a diagonal matrix which is rectangular. The magnitude of diagonal values of matrix W are organized in a descendant order. Such diagonal values are identified as singular values of matrix C. For matrix C with m×m order, Singular matrix (W) can comprise of utmost m diagonal values. Layer wise involvement of L and R matrices in the matrix C is represented through singular values; over which the fragmentation is accomplished. The matrix quality becomes worse when the order of regeneration of matrix decreases . In contrast, there is no much disturbance in the quality of matrix with a minor disparity in the singular values. Eq. 2 shows one of the most important properties of singular values of diagonal matrix W.

$$f_1 \geq f_2 \geq f_3 \geq \ldots\ldots\ldots > f_{k+1} > d_{k+2} = d_m \quad (2)$$

Where 'k' , ($k \leq m$) indicates the rank concerned to singular matrix W, $f_1, f_2, f_3, \ldots$ f_m symbolizes the diagonal values and m denotes the order of matrix. The matrices L and R likewise follow: $LL^T = I_m$ and $RR^T = I_m$.

2.2. Discrete wavelet transform (DWT)

The knowledge of wavelets is exploited by DWT to depict the images in an energy compact manner [17]. Wavelets are delineated as the small waves with frequencies which are always changing. DWT offers a great energy packaging capacity and explicitly it is the reason to apply in numerous image processing needs such as compression of the image, watermarking of the image, etc. In DWT, the novel image is changed into wavelet domain, consisting of four bands. Among the four sub-bands, three (HL, LH and HH) are identified as sub-images which encompass high frequencies with fringe data. The other sub-band (LL) comprises of low frequency, termed as approximate sub-image since it holds an estimated data of original image. It is attained by low-pass filtering together in the row as well as column directions. Additionally decomposition of sub-bands is feasible to reach higher level of decomposition. This progression can be passed to numerous levels as required in the application. When related to HL, LH and HH sub-bands, LL sub-band proves to be more stable in nature and it is extremely tough to get transformed by attacks as this band holds principal portion of energy of original image. This feature makes the approximate sub-band more apt for concealing the watermark with robustness. The exceptional spatio-frequency localization property pertained to DWT is exploited by disparate image watermarking methods in recent to contribute a raised value with regard to imperceptibility besides robustness. Figure 1 depicts disparate levels concerned to discrete wavelet transform.

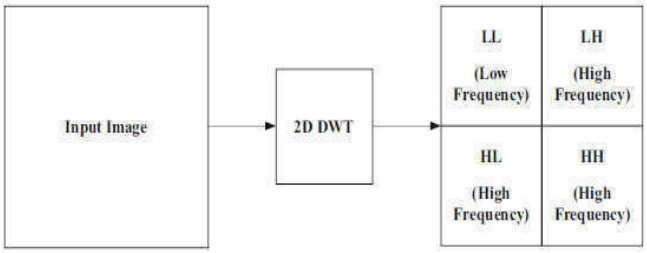


Fig. 1 First level decomposition of wavelet for an image

2.3. Artificial bee colony (ABC)

Among several evolutionary procedures [18] identified Artificial Bee Colony (ABC) as one such simple, rapid population based method and performance wise also it is too robust. ABC is practiced effectively on disparate domains to offer relatively raise in results [19-22]. Analogous to other Evolutionary procedures, variable values which are optimal are to be inspected by exploiting the ABC method with the aim of maximizing/ minimizing the specified cost function. Simulation outcomes demonstrates that ABC attains identical results comparable to several dissimilar metaheuristics practices along with this it also presents an extra benefit of adopting fewer control constraints. ABC is familiar for its capability to explore the search related to multidimensional which is the main focus of the projected work also. The basic steps intended for implementation of ABC algorithm can be realized from [23].

ABC method functions according to the sharp foraging behavior of the honey bee swarms. The feasible solutions for any specified issue is categorized by the food source in ABC scheme, further the fitness of any outcome is measured by the quantity of nectar in a food source. Three forms of bees arise in ABC, termed as working bees, spectator bees and detective bees. Working bees directs the number of explanations in the

specified population dimensions. ABC is originated with an initial population of results of dimension L (places where food is available) with each having a dimension X. i.e., early solution can be symbolized as P_k={p(k,1), p(k,2)……p(k,X)}; Where k=1,2…..L.

The searching process is a repetitive procedure subsequent to the initial circulation, while selecting optimum solution until the ending state is obtained. The spectator bees opt for the sources of food which are finest, depending on the value of fitness function in addition to the information provided by working bees, while the detective bees kept their recent sources of food with the purpose of exploring improved sources of food. In ABC procedure, working bees and spectator bees are answerable for the process of exploitation, while detective bees take attention of appropriate inspection.

The ABC algorithm encompass the following stages:

1. The sources of food with L number are preferred arbitrarily among the acceptable lower P_{min} = ($p_{(min,1),\ldots\ldots}$ $p_{(min,X)}$) as well as upper limit P_{max} = ($p_{(min,1),\ldots\ldots}$ $p_{(min,X)}$) of distribution with each including dimension of X through Eq.(3)

$$p_{(k,l)} = p_{(min,l)} + rand(0,1) \times (p_{(max,l)} - p_{(min,l)}) \quad (3)$$

2. All working bees produces a novel result with the local information accessible to it and relates the fitness of engendered result with the parent result. The improved result between these two is utilized for succeeding iteration. The novel result Q_k is engendered from existing result by making use of Eq.(4)

$$q_{(k,l)} = q_{(k,l)} + ¢_{(k,l)} \times (q_{(k,l)} - p_{(m,l)}) \quad (4)$$

Here $¢_{(k,l)}$ is an arbitrarily acquired numeral between [-1,1] and indices m € (1,2,….L) and l € (1,2,……X) are arbitrarily acquired in such a form that m and l persist to be unlike.

3. The working bee distributes the fitness details with the spectator bee. The spectator bee produces a probability (Pi) of nectar (fitness) quantity by means of Eqs. (5) and (6).

$$F_K = \frac{fit_K}{\sum_{k=1}^{L} fit_K} \quad (5)$$

$$fit_k = \begin{cases} \dfrac{1}{f(Y_k)+1}, iff\,(Y_k) \geq 0 \\ 1 + abs[f(Y_k)], otherwise \end{cases} \quad (6)$$

At this point $f(Y_k)$ denotes the value of objective function at the position of food Y_k.

4. An arbitrary numeral is engendered for all spectator bees within zero and one; if Fk , the position of food possess larger value than the arbitrary numeral then step 2 is tracked by the spectator bee as well.
5. If prearranged number of repetitions cannot alter the positions of food, at that time such positions are supposed to be eliminated. The value of prearranged numeral of repetitions is considered as significant constraint and identified as boundary. In such circumstances, detective bees regulates the novel places arbitrarily with the aim of exchanging the places that are eliminated.

   Stages 1 to 5 are reproduced up to the prearranged ending condition (maximum repetition, least variation) encounter.

3. PROPOSED ABC BASED DWT-SVD VIDEO WATERMARKING TECHNIQUE

3.1. Watermark Embedding

The planned watermarking method comprises primarily of two phases; watermark concealing and watermark extortion. For embedding watermark into video, we prefer YUV QCIF videos and partition the video clips into video scenes. The frames of video scene are evaluated. The frames of YUV video are in RGB color model and they are transformed to YCbCr color model. The luminance part is not far delicate to amendments when analogized with chrominance elements (Cb & Cr). For that reason while implanting watermark, take away luminance element. The procedure of embedding and extracting the watermark is shown in Figures 2 &3 respectively. Sequentially to entrench the watermark into all frames the consequent steps are taken.

Step1. The original video 'V' is segregated as set of M number of frames, the watermark 'W_m' is entrenched inside the video 'V'.

Step2. In the group each frame is transformed into YC_bC_r color model from RGB, further the luminance values Y of the frames are later accomplished.

Step3. In the video 'V' only luminance frames are alienated into 8x8 non-overlapping blocks.

Step4. Apply 2D DWT on entire 8x8 luminance frame blocks of the video 'V' as shown in Eq(7).

$$V_{DWT} = DWT\,(V) \quad (7)$$

Step5. ABC method is employed primarily to elect particular 8x8 blocks signified as Sel_dwtblock to embed watermark in every frame for attaining perceptibility.

Step6. SVD transform is implemented on the Sel_dwtblock of V_{DWT} of all frames to obtain left singular matrix L and singular matrix W as well as right singular matrix R. The eigen values are at diagonal location of each block. Majority

of the strength is preserved in the eigen values of the video, besides we attain Lv in addition to Wv, as depicted in Eq(8).

$$Cv = Lvb \times Wvb \times Rvb^T \quad (8)$$

Where b represents the number of the blocks selected for concealing the watermark.

Step7. SVD transform is applied to the watermark and we obtain L_{wat} and R_{wat}, as shown in Eq(9)

$$Wat = L_{bwat} \times W_{bwat} \times R_{bwat}^T \quad (9)$$

Step8. We add Wbwat to the Wvb matrix to obtain W_{vbnew} as given in Eq(10)

$$W_{vbnew} = Wvb + \beta Wbwat \quad (10)$$

The selection of the 'β 'coefficient depends on the compromise strength or imperceptibility preferred by the user.

Step 9. Inverse SVD is applied to obtain modified selected blocks as given in Eq(11).

$$C_{vnew}=L_{vb}\, W_{vbnew}\, R_{bv}^T \quad (11)$$

Step 10. Reposition the modified blocks in the frame of video to obtain V_{DWTnew}

S*tep11.* Apply Inverse DWT which is defined in the following Eq. 12

$$R_{wat} = DWT^{-1}(V_{DWTnew}) \quad (12)$$

3.2. Watermark Extortion

The gradual progression for the extortion of watermark in detail is delineated as follows:

Step1. The attained watermarked video is separated into clusters of k frames.

Step2. Every frame of the cluster is reformed from the RGB into the YCbCr color model.

Step3. Each luminance frame is transformed into 2D DWT.

Step4. Segment all luminance frames of the video 'V' as non-overlapping 8x8 blocks.

Step5. Spread on 2D DWT on entire 8x8 blocks of the luminance frame of the video, 'V' and the dwt block with outsized texture value is extorted which is signified as Sel_dwtblock by exploiting ABC method where watermark is entrenched.

Step6. Simulate SVD transform on all the Sel_dwtblock of V_{DWT} of each frame and obtain L, W and matrix R as follows:

$$C_v' = L_{bwat}' \times W_{bwat}' \times R_{bwat}^T \quad (13)$$

Step7. The diagonal entries are Eigen values of every block and are obtained by the following Eq (14)

$$Wbext = \frac{W_{vbnew} - Wvb}{\beta} \quad (14)$$

*Step 8.*The watermark is rebuild using the following Eq (15)

$$WE_{EXT} = L_{wat}\, W_{wat}\, R_{wat}' \quad (15)$$

Where WE_{EXT} is the extracted watermark

4. SIMULATION RESULTS AND ANALYSIS

The entire simulation results are assessed by finding the imperceptibility and robustness. The simulation results with proposed algorithm are contrasted with those of [23] and [24] for testing primarily two criteria's: imperceptibility test as well as robustness test. The superiority is shown in two evaluations; they are: Peak Signal to Noise Ratio (PSNR) on top of Normalized Correlation (NC) for mutually before and after attacks.

4.1 For Testing Imperceptibility

The simulated value of NC is 1 i.e. the extorted watermark is alike to the original and precise extortion is accomplished. While relating the planned scheme with [23] and [24] the projected scheme recorded a PSNR up to 59.93db for Suzie video which is superior to the PSNR stated by [23] and [24]. Table.1 demonstrates the PSNR assessment of the projected video watermarking procedure with earlier hybrid optimized video watermarking methods. The performance evaluations of the planned method with former hybrid optimized video watermarking schemes are shown in Table.2 for the Suzie video sequence in terms of the NC.

Table.1. Relative PSNR assessments with [23] and [24] after dissimilar sorts of attacks on the video of Suzie

Different types of Attacks	PSNR		
	Proposed method	[23]	[24]
(a) No attack	59.93	48.12	58.45
(b) Resizing	54.22	48.98	52.34
(c) Rotation	57.63	49.54	54.67
(d) Cropping	53.58	39.99	50.78
(e)Frame swapping	54.44	47.34	51.16
(f) Sharpening	56.32	38.67	50.72
(g)Frame Averaging	54.45	44.54	53.69
(h)Salt &Pepper noise	54.89	47.86	49.25
(i)Gaussian noise	54.12	43.97	52.02
(j)Histogram Equalization	53.04	39.35	50.30

ISBN 978-81-939386-1-4

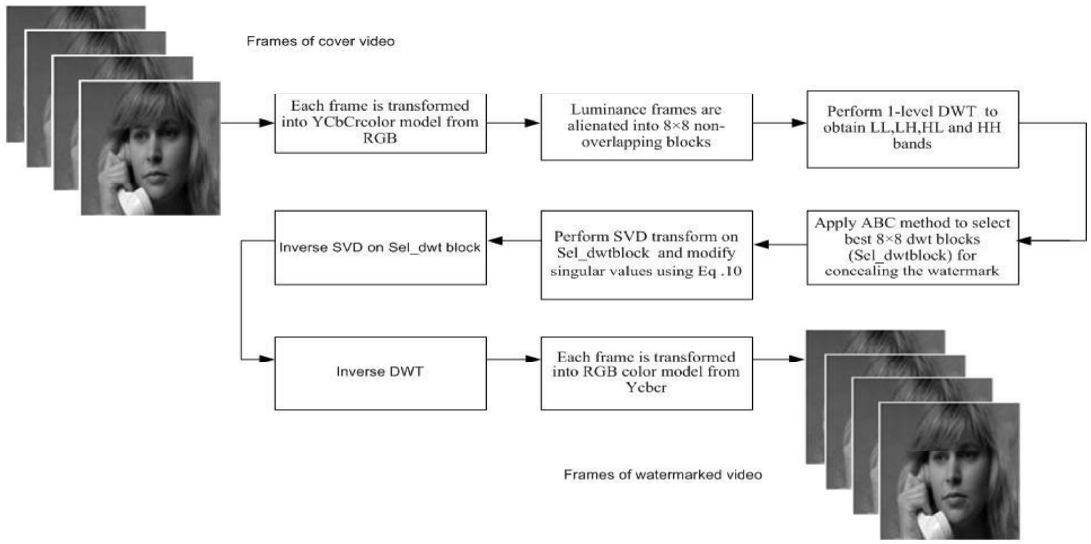


Fig.2.Procedure of Embedding the watermark

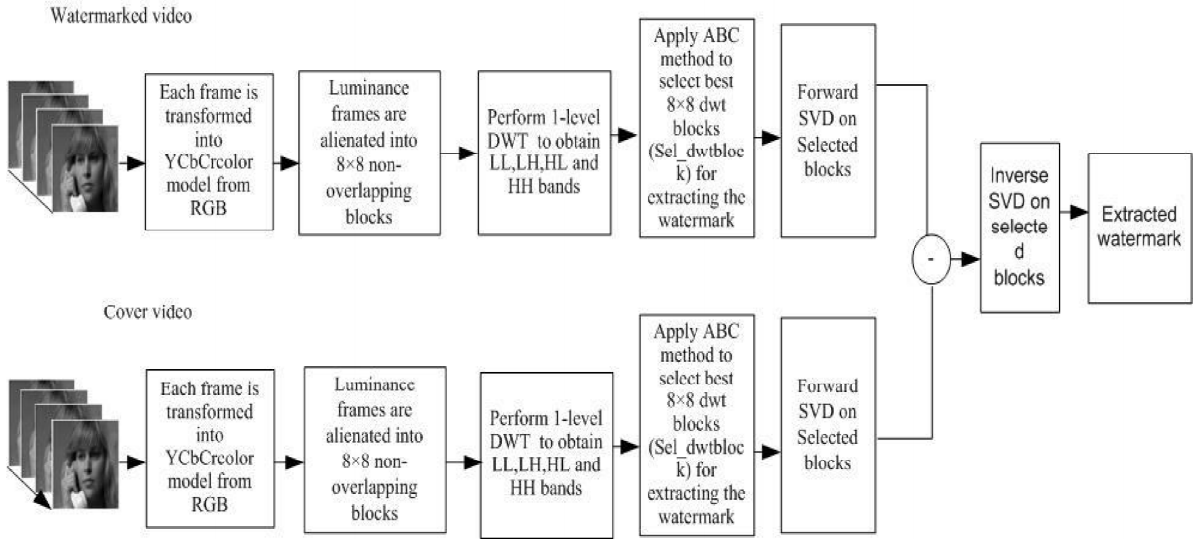


Fig.3.Procedure of Extracting the watermark

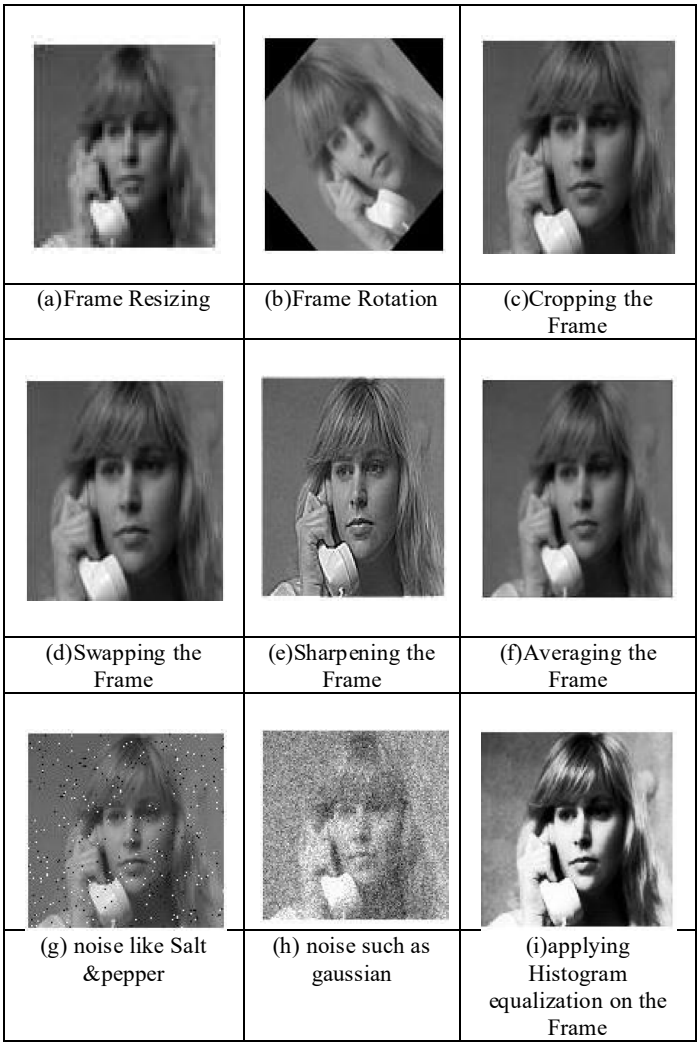
Table.2. Relative NC values with [23] and [24] after dissimilar sorts of attacks on the video of Suzie

Different types of Attacks	**Proposed scheme**	**[23]**	**[24]**
(a) No attack	1	0.9	1
(b) Resizing	0.99	0.84	0.93
(c) Rotation	0.95	0.89	0.9
(d) Cropping	0.98	0.89	0.91
(e)Frame swapping	0.88	0.7	0.8
(f) Sharpening	0.93	0.76	0.87
(g)Frame Averaging	0.91	0.79	0.91
(h)Salt &Pepper noise	0.99	0.87	0.92
(i)Gaussian noise	0.9	0.85	0.86
(j)Histogram Equalization	0.89	0.6	0.82

4.2. For Testing Robustness

Robustness is the capability of detector to extort the masked watermark even after altering the watermarked data. Owing to enormous volumes of data and intrinsic redundancy between frames, the video used for embedding the watermark is extremely susceptible to plagiarize assaults which include

averaging of frames, rotation, tumbling of frames, sharpening etc. Frequently robustness is assessed by the endurance of watermark after disparate varieties of attacks, such as Resizing of frames, Rotation of frames, Cropping the frame, swapping the frame, Sharpening the frame, averaging the frame, effect of noise like Salt & pepper besides Gaussian and applying Histogram equalization on the frames, etc. The retrieved watermark after applying different types of attacks as shown in Fig.4. on Suzie video is depicted in Fig.5.

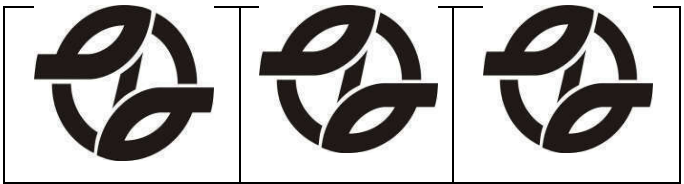


(a)Frame Resizing | (b)Frame Rotation | (c)Cropping the Frame

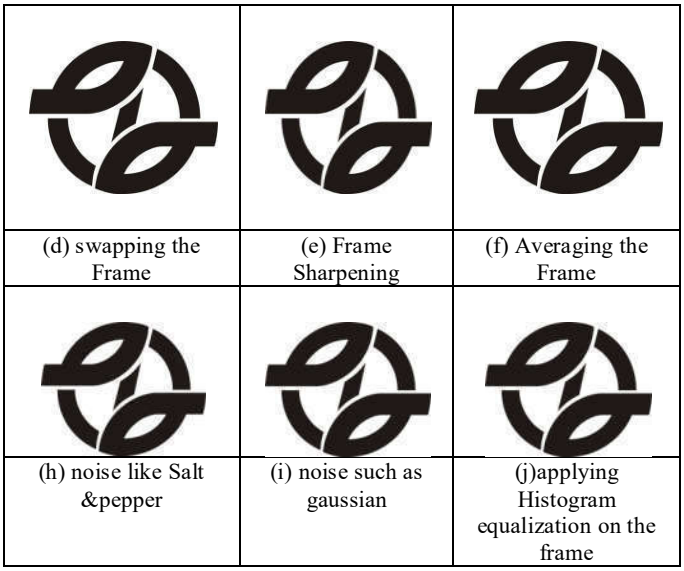
(d)Swapping the Frame | (e)Sharpening the Frame | (f)Averaging the Frame

(g) noise like Salt &pepper | (h) noise such as gaussian | (i)applying Histogram equalization on the Frame

Fig.4. Disparate varieties of attacks on the video of Suzie



(a)Frame Resizing | (b) Frame Rotation | (c) Frame Cropping



(d) swapping the Frame | (e) Frame Sharpening | (f) Averaging the Frame

(h) noise like Salt &pepper | (i) noise such as gaussian | (j)applying Histogram equalization on the frame

Fig.5. Extracted watermark after applying dissimilar types of attacks on the video of Suzie.

In the proposed scheme the NC values are always above 0.85, revealing the developed quality of the extorted watermark. Thus the recommended algorithm illustrates upright robustness though the attack is resilient. The object next to the greater concert of the planned algorithm beneath numerous kinds of attacks is as a result of: relating superior categories of transforms in union to optimization procedure, therefore limited numerals of coefficients interrupt on the excellence of the restored watermark; also gradually it authorizes us to exploit highest entrenching feature with enhancement in robustness.

5. CONCLUSION

In this work, an optimized DWT–SVD block-based video watermarking scheme is presented. Primarily, the video is segregated into blocks. Subsequently, DWT is applied on all of these blocks then only selected blocks were preferred by ABC method to attain high-level rating for the watermarking constraints and sustain the trade-off among them, finally the SVD characteristics were exploited to choose the best embedding regions to include the watermark, to attain more robustness. The embedding strength applied at the period of inserting watermark is so principal that only the appropriate owner can recover the watermark, identifying the legitimate constraints. The projected scheme outperformed all related earlier schemes and accomplished the highest imperceptibility. In terms of robustness, good resiliency was examined in opposition to all types of video processing attacks and numerous types of geometrical attacks.

REFERENCES

[1] Miller, M. L.; Cox, I. J. & J. A. Bloom. (2002). Digital Watermarking. Academic Press, 2002.

[2] Koch E.; Zhao, J. & Luo, C. (1998). In business today and tomorrow Communications of the ACM, 41, July 1998. pp 66–72.

[3] K.Swaraja, Protection of Medical Image Watermarking, Journal of Advanced Research in Dynamical and Control Systems (JARDCS), Special Issue 11, July 2017, ISSN: 1943

[4] T. Jayamalar, V. Radha, "Survey on digital video watermarking techniques and attacks on watermarks", International Journal of Engineering Science and Technology: Vol. 2(12), pp. 6963-6967, 2010.

[5].Swaraja, K., Latha, Y.M., Reddy, V.S.K. and Paramkusam, A.V., 2011, December. Video watermarking based on motion vectors of H. 264. In India Conference (INDICON), 2011 Annual IEEE (pp. 1-4).

[6]K.Swaraja, Y.Madhaveelatha, V.S.K.Reddy, "A Pristine Digital Video Watermarking in H.264 compressed domain", IEEE International conference on computational intelligence and computing research (ICCIC), Dec 18-20, 2014, pp.1-4, Coimbatore, India, ISBN: 978-1-4799-3974-9.

[7] Swaraja, K., Madhaveelatha, Y. and Reddy, V.S.K., 2014, December. A pristine digital video watermarking in H. 264 compressed domain. In Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on (pp. 1-4).

[8] Ashish M. Kothari, Ved Vyas Dwivedi, "Performance Analysis of Digital Video watermarking using Discrete Cosine Transform", International Journal of electrical and Computer engineering systems, Vol. 2, Number 1, 2011.

[9].Meenakshi, K., Prasad, K. S., & Rao, C. S. (2017). Development of Low Complexity Video Watermarking With Conjugate Symmetric Sequency– Complex Hadamard Transform. IEEE Communications Letters, 21(8), 17791782. 14.

[10] Nisreen I. Yassin, Nancy M. Salem, and Mohamed I. El Adawy, "Block Based Video Watermarking Scheme Using Wavelet Transform and Principle Component Analysis IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012.

[11].Meenakshi, K., Ch Srinivasa Rao, and K. Satya Prasad. "A robust watermarking scheme based Walsh-Hadamard transform and SVD using ZIG ZAG scanning." In Information Technology (ICIT), 2014 International Conference on, pp. 167-172. IEEE, 2014.

[12] B. Chandra Mohan and K. Veera Swamy Z, " On the use of Schur Decomposition for Copyright Protection of Digital Images", International Journal of Computer and Electrical Engineering, Vol. 2, No. 4, August, 2010,pp. 1793-8163.

[13] Swaraja, K. "Medical image region based watermarking for secured telemedicine." Multimedia Tools and Applications 77 (2018): 28249-28280.

[14] K.Swaraja, Y.Madhaveelatha, V.S.K.Reddy, "A secure method of optimized low complexity video watermarking", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 4, MARCH 2015, ISSN 1819-6608, pp.1822-1827.

[15] Ansari, I.A., Pant, M., 2015. SVD watermarking: particle swarm optimization of scaling factors to increase the quality of watermark. In: Proceedings of Fourth International Conference on Soft Computing for Problem Solving, Springer India, pp. 205–214.

[16] Meenakshi, K., Srinivasa Rao, C. and Satya Prasad, K., 2014. A scene based video watermarking using slant transform. IETE Journal of Research, 60(4), pp.276-287.

[17] Burrus, C. S., Gopinath, R. A., & Guo, H. (1997). Introduction to wavelets and wavelet transforms: a primer.

[18] Karaboga, D., Akay, B., 2009. A comparative study of Artificial Bee Colony algorithm. Appl. Math. Comput. 214, 108–132.

[19] Ansari, I. A., Pant, M., & Ahn, C. W. (2016c). ABC optimized secured image watermark-ing scheme to find out the rightful ownership. OptikInternational Journal for Light and Electron Optics, 127(14), 5711-5721.

[20] Draa, A., Bouaziz, A., 2014. An artificial bee colony algorithm for image contrast enhancement. Swarm Evolut. Comput. 16, 69–84.

[21]. Akay, B. (2013). A study on particle swarm optimization and artificial bee colony algorithms for multilevel thresholding. Applied Soft Computing, 13(6), 3066-3091.

[22]. Hanbay, K., Talu, M.F., 2014. Segmentation of SAR images using improved artificial bee colony algorithm and neutrosophic set. Appl. Soft Comput. 21, 433–443.

[23]Priyanka Singh, and Suneeta Agarwal, "A Hybrid DCT-SVD Based Robust watermarking scheme for copyright protection", International Conference on Emerging Trends in Engineering and Technology (ICETET'2013) Dec. 7-8, 2013 Patong Beach, Phuket (Thailand),pp.1-5.

[24]. Ansari, I. A., Pant, M., & Ahn, C. W. (2016d). Artificial bee colony optimized robust-reversible image watermarking. Multimedia Tools and Applications, 1-25.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Custom Queueing method for enhancing security of a Packet Scheduling Process in Real time Networks

1st G.Karuna
Computer Science and Engineering, GRIET
Hyderabad, Telangana, India
karuna.g@griet.ac.in

2nd Swaraja K
Electronics and communication Engineering, GRIET Hyderabad, Telangana, India

3rd Meenakshi.K
Electronics and communication Engineering, GRIET Hyderabad, Telangana, India

4th Padmavathi.Kora
Electronics and communication Engineering, GRIET Hyderabad, Telangana, India

***Abstract* — The wireless technologies became a part of our day to day life, for sending and sharing data very rapidly so that everyone today uses wireless networks for their business or personal. The traditional wireless communication technology is unable to satisfy real-time transmission requirements in mobile electronic commerce application such as access of books from digital libraries, access of stock exchange information etc. Recent real-time wireless networks require high level security to guarantee the information passed through wireless channels. So, designing flexible security mechanisms for real time applications and transmitting packets through wireless networks is highly desirable. The present work introduces a new packet scheduling algorithm to enhance security which mainly aims to dynamically determine security levels of packets based on security requirements of an application and also guaranteeing deadlines for packets. By incorporating Custom Queuing method with this packet scheduling algorithm, the overall performance of real time networks is increased.**

Keywords— Wireless Networks, Packet Scheduling, Custom Queuing

I. INTRODUCTION

Wireless technology today has become one of the most affordable and easy to use in the field of information technology and academics because of its fastest growth of various applications in recent are using wireless networks. Almost every common man today can access to the Internet at their homes, libraries, universities, restaurants, railway stations, airports, and even while driving in vehicles to conduct their business, accessing and learning e-books, transfer money, paying electricity bills, booking movie tickets, sharing stock exchange information etc.

As Internet users become increasingly mobile and online business applications become very much interactive and traditional wireless communication technology methods are unable to satisfy the real-time transmission requirements in mobile electronic commerce applications such as access of books from digital libraries, access of stock exchange information etc. The remedy to this problem is real-time wireless communication techniques allowing users to collect and transmit data in a timely manner increases interest in many academicians, scholars and researchers. Supporting efficient and reliable data transmission, especially real time data transmission, through wireless networks is extremely difficult and challenging task because wireless networks must be facing more complicated environments compared with conventional wired networks such as computer viruses, intruders, worms, spy wares, and similar threats.

Many security policies related to authenticity and confidentiality strategies and wireless communication protocol-based security schemes [1] have been proposed and applied in real-time wireless networks. However, most of them considered only security issues in a static mode, in which security levels are all configured when wireless network systems are built. In some real-time systems like stock quote updating and trading system, users may need flexible quality of security, measured as security levels. Suppose in an industry the data may be more secured when compared to the past twenty years data in that company. i.e security levels are higher at present for that company. Thus designing flexible security mechanisms for real time applications transmitting packets through wireless networks is highly desirable.

The Custom Queuing method for improvement of Security Aware Packet Scheduling Algorithm aims to dynamically determine security levels of packets according to applications, security requirements while guaranteeing deadlines for packet and also increases the speed i.e overall performance. Apart from achieving high quality of security, it can significantly improve guarantee ratio, which is a fraction of total transmitted packets that are found to be delivered before their deadlines.

II. RELATED WORK

There is a rapid growth of applications in the domain of wireless networks in the real world, i.e. usage of Internet at all places in homes as well as business for sales, purchases, finding remote locations, making payments and playing games and many more. Because enormous development of Wireless technologies makes a person day to day life more comfortable. Many devices with wireless technology developed and used for an individual and for business point of view which leads an increase of users every day. Security risks automatically raised and it is a challenging task for secured transmission of data. National Telecommunications Cooperative Association (NTCA) reported in the year of 2005 nearly 61% of survey

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

defendants provided wireless services to their regular customers and nearly 55% of real time services are for voice like services. In the coming years Wireless technologies will offer further many new features and functions for protecting data and safe transmission.

Wireless technologies, makes easy for communication due to no cables required for connection. i.e. two or more devices are enabling communication without connecting physically each other with transmission cables. Instead they use micro and radio waves [2] with specific range of frequencies for providing communications among larger distances. Depending on speed, coverage area and size there are many categories of wireless networks available. Wireless personal area networks which covers very smaller distances within a room (Bluetooth), Wireless Local Area Networks that covers few kilometers by connecting laptops with the device having wi-fi access point, wireless metropolitan area networks range hundreds of kilometers with multiple nodes and wireless wide area networks covers the distance across worldwide through various cellular technologies (WiMAX, CDMA, GSM etc) [3]. Best example for Wireless Wide Area Network is an Internet. Wireless technologies range from complex systems to simple hand held devices [4]. Different packet scheduling algorithms surveyed from exited work [5, 6].

A. Security Issues

Security plays an important role in the design and development of wireless mobile educational and business applications, international wireless organizations, wireless equipment providers and thus academic researchers made extreme efforts in increasing the features of existing security mechanisms [7, 8] and finding attractive and creative security policies of wireless networks. IEEE introduced many standards for improving security by defining various security related functions. Cisco provides the solutions for wireless applications by using strong encryption technology and providing unified WLAN. Papers addressing the security problems also provide valuable solutions for wireless business applications.

In recent years, wireless networking has become more popular and users are accessing wireless technology while moving also for their daily activities. Thus, protecting data from unauthorized access or the attacks like active and passive attacks etc is a challenging issue apart from efficient and reliable transfer of data. There are many existing algorithms to improve performance of wireless networks. This paper highlights those security threats, and explains what we need to know to use wireless safely, both in the home and in public.

B. Packet Scheduling

Packet Scheduling methods are generally used when multiple packets exist in a buffer and they share a common outgoing link. Packet scheduling algorithm determines the service policy of a node and it plays important role for providing quality of service. Many packet-scheduling algorithms exist in the literature for increasing the quality of service within a node while delivering packets from source node to destination node. Chang and Yu presented packet-scheduling algorithms [9] to guarantee the quality of service in wireless applications of ATM and video traffic. Different packet scheduling algorithms developed for achieving quality of service in the nodes such as FIFO, Priority Queue, Weighted fair queue and low latency queue etc. In FIFO method the packets are served based on their arrival order. Priority Queues gives preference for important traffic to be cleared during transmission. Weighted fair queue works based on weighted bandwidth allocation. It divides bandwidth across multiple queues. However most of these packet scheduling algorithms works well for better delivery packets and achieve quality of service, but not concentrating on security constraints. This paper mainly integrates the proposed packet-scheduling algorithm [10,11] with dynamic security adjustment strategy. In doing so, we build a new secure packet scheduling scheme along with custom queueing policy for real-time wireless networks.

III. PACKET SCHEDULING ALGORITHM WITH SECURITY CONSTRAINTS

The packet scheduling Algorithm with security has been introduced with the aim to increase the performance and reliable delivery of packets by using various security levels [14]. This Security based Packet scheduling algorithm improves the performance by raising security levels where a security level controller is used.

Initially packets are sent into the accepted queue and security levels are increased for every packet at accepted queue by determining the guarantee and deadline of packets. When packet has considerable deadline, it is admitted into accepted queue otherwise it will be sent into the rejected queue. The below given constraint verifies whether packet has considerable deadline.

$$CTi - STi <= di$$

where,
STi is the start time of transmission of the ith packet,
CTi is the transmission completion time, di is the deadline of packet.

Based on the deadlines of the packets, they are processed. i.e. the packet which is arrived first is scheduled first. This algorithm is initiated with very low-level security levels. Further, the security is increased based on below constraints.
(i) when Pi is transmitted earlier than given deadline.
(ii) Pi packets are guaranteed even the deadlines have been inititaed later.

The above constraint (ii) is important and reasonable if the packet is arrived in real time link, then its timing constraint has to guaranty the arrival. Simply this security algorithm confirms that an admitted packet is not harmfully affected by simultaneous admitted packets.

A. The Custom Queuing for Efficiency

This strategy works for fixed-length packets scheduling [10,11]. FIFO within priority queues. It is based on interface or protocol. Time is divided into frames. Each session is assigned a weight in terms of number of packets that could be served during a frame. For every session a separate counter is kept in order to register the number of packets served in that specific frame. It guarantees bandwidth per queue. It is not used for voice. It increases speed and performance during transmission of packets.

B. Algorithm

//calculate no. of packets to be served each round//
min = find smallest weight
For each flow f
f.packets_to_be_served=f.weight/min
//main loop
Loop
For each non_empty flow queue f
min(f.packets_to_be_served,f.packets_waiting).times
do
serve packet f.get packet

IV. SYSTEM MODEL

The system model includes a wireless channel as an NN switch [2] shown in Fig.1. A transceiver is used for transmission of packets even though a separate transmitter and receiver is equipped by each wireless node in the system. As such, a node cannot transmit and receive packages simultaneously. A packet scheduler matching transmitter is equipped to the corresponding receivers in this model.

In the above-mentioned model, a centralized scheduler collects, via a control channel, global network information. The scheduler uses a slotted system and a matching algorithm 2 to match transmitters to their intended receivers. Clearly, there is no need for a transmitter to transmit to the receiver on the same mobile node. Therefore, the possibility of a transmitter sending to the receiver on the same mobile node is disallowed. The resulting system of N transmitters and N receivers, with exactly one transceiver on each mobile node, can be modeled as a special case of an NN switch. The N transmitters can be thought of as the N input ports of a switch and the N receivers as the N output ports. The centralized scheduler node is modeled as the arbiter or scheduler of the switch. This PAN model assumes that each node's transmission can reach any of the other N_i nodes and the central node3.

A matching in a bipartite graph G = (V,U,E) is a subset of edges G (i.e., of E) such that no node in V or in U has more than one edge of this subset incident on it. A maximal matching in a bipartite graph is any matching such that given this matching, no other edge can be added to it without violating the definition of the matching. A maximum matching in a bipartite graph G is a matching consisting of the maximum possible number of edges in G. The central node or the scheduler could be either one of the N nodes or a separate entity.

This also includes a PAN, based on this model a Wireless Switched PAN or simply a Switched PAN. With these assumptions, the model can be represented by a special bipartite graph.

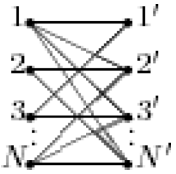


Fig 1: N-Node Bipartite Graph (NNBG)

There are three more components in the system apart from switch model, including Security Controller (SC), an Admission Queue Controller (AQC), and Deadline Controller (DC) as depicted in Fig. 2.

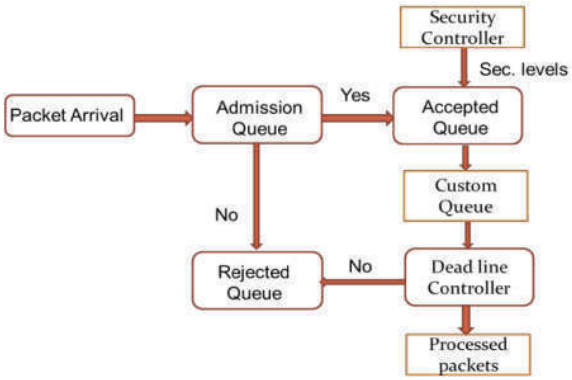


Fig.2. System architecture

This model represents a connection between two nodes in specified wireless network. And the packets are submitted independently to the wireless link with arrival rates abided by Poisson distribution. The Admission Queue Controller mainly decides given packet is accepted or not. If it is accepted then sent into accepted queue otherwise into rejected queue. The security levels for packets in the accepted queue will be assigned by the Security Controller. The Deadline Controller scheduler uses deadline policy for all admitted packets in which security levels have been increased by the Security Controller.

A. The Packet Model

The packets transmitted in this model have deadlines to reach destination node and every packet to be transmitted individually. All the packets have arrival times based on the classical Poisson distribution. Packet Pi is represented as a tuple (ATi, PTi, DLi, SLi), where ATi and PTi represents the arrival time and the processing time of packet i. DLi depicts deadline and SLi specifies the security level of packet *i*. Besides, without loss of generality we assume that each packet is assigned a quality of security measured as a security level *SLi* that in the range [1, 2, ..., 10], where 1 and 10 are the smallest and largest security levels. The security level 1 means that smallest security level of the packet and level 10 is largest. To calculate the security overhead without loss of generality, the below given formula (1) is used to model the security overhead envisioned as the extra processing time experienced by packet *i*.

$$SOi = ETi * (SLi / R) - (1)$$

where SOi and SLi depicts security overhead and security level of packet i, ETi is the transmission time of the packet and R is set to 10. Thus, the total processing time WLi of packet i can be expressed as:

$$WLi = ETi + SOi = ETi * (1 + SLi / R)$$

Where, SOi-security overhead of packet
SLi - security level
ETi- transmission time of packet
R value set to 10

V. SIMULATION AND RESULT ANALYSIS

The transmission of packets with in a node while delivering from source node to destination node using custom queuing method is shown in the given table 1.

The simulation parameters considered for SPSCQ algorithm are 60 source nodes, 40 destination nodes and 5 intermediate nodes. The packets are transmitted from source node to destination node through intermediate nodes by changing security levels ranging 1-10. The arrival rate is 5 packets/sec. bandwidth taken as 1mbps.

(i) Guarantee % = (no.of accepted packets/no.of arrived packets) *100
(ii) Agg. security % = (security level used for accepted packets/ no. of arrived) *100

Table 1. Packet schedule process with CQ

P. No.	Generation of packets				Acc/ rej	Incr. sec	With Custom Queuing (CQ) Method (packets)
	St (s)	End (s)	dead line	Secu. level			
1	1	3	15	2	acc	7	[4,4,9,10, 7]
2	2	5	14	6	acc	6	Parallely sending [5,5,11,10 ,9]
3	3	7	17	8	acc	8	[2,2,5,14, 6]
4	4	9	10	7	acc	7	Parallely sending [10,10,20, 14, 7]
5	5	11	10	4	acc	9	[1,1,3,15, 7]
6	6	13	4	1	rej		---
7	7	15	18	3	acc	8	[3,3,7,17, 8]

The guarantee percentage (GP) and aggregate security level percentage (ASP) of SPACQ is calculated and compared with other algorithms SPSS [12,13], ISAPS[15], RSAPS[16] as shown in above fig2 and fig3.

The experimental results shown that the new SPSCQ approach provides an improvement in terms of guarantee ratio, security level, and speed. Thus, the overall performance for real time wireless networks has been improved using Custom Queuing approach and the security levels organized dynamically based on the network traffic.

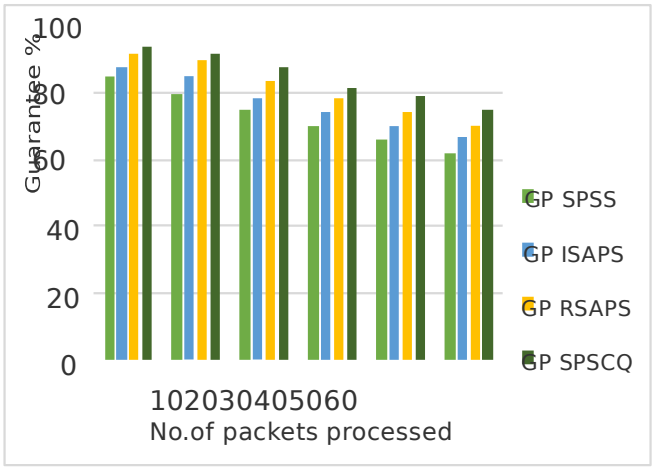


Fig2. Guarantee % (GP) of various Algorithms

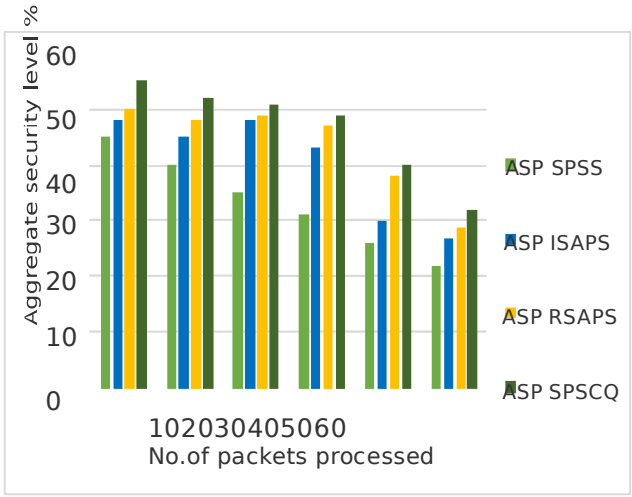


Fig3. Security % (ASP) of various Algorithms

VI. CONCLUSION

In modern wireless networks high quality of security is essential apart from guarantee ratio of packets, in order to protect data which is stored in the packets to be transmitted. The proposed work suggested a novel dynamic Security based Packet Scheduling algorithm with Custom Queuing approach (SPSCQ), which is capable of achieving high quality of security service for real-time packets and reliable while delivering packets. The algorithm is also designed in a way that makes it possible to achieve a reasonably high guarantee ratio and optimized security level. In particular, the proposed SPSCQ algorithm leverages an intelligent Security Level Controller to adaptively assign security levels to incoming real-time packets transmitted through a wireless network links. The performance of the present approach can be further improved by using combinations of various other packet scheduling algorithms.

REFERENCES

[1] http://www.ntca.org/ content documents/ 2005 WirelessSurveyReport.pdf

[2] S. Al-Harthi and R. Rao, "A switch model for improving throughput and power fairness in Bluetooth piconets," in Proc. Globecom 2003, pp. 1279–1283.

[3] http://www.wifinotes.com/wireless-networks.html

[4] T. Karygiannis and L. Owens, "Wireless Network Security 802.11, Bluetooth and Handheld Devices," chapter 2, http://csrc.nist.gov/publications/nistpubs/ 800-48/NIST SP 800-48.pdf

[5] M. A Badamas, "Mobile computing systems–security considerations," Information Management and Security 2001, pp. 134–136.

[6] W. Kensaku, K. Shoji, T. Yutaka, K. Yoshinobu, I.Eisaburo, A packet scheduling algorithm for max-min fairness in multihop wireless LANs, Comput. Commun. 32 (2009) 1437–1444.

[7] Z. Zhang, S. Bronson, A packet scheduling algorithm for optimizing downlink throughput in wireless LANs with the one-sender–multiple-receiver technique. in: Proc. the 28th IEEE Conf. Global Telecommunications (GLOBECOM 2009), Dec. 2009, pp. 1–6.

[8] V. Gupta and S. Gupta, "Securing the wireless Internet," IEEE Commun. Mag., pp. 68–74, Dec. 2001.

[9] C.Chang J.Chang and M.Yu "Guaranteed Quality of Service wireless access to ATM".IEEE J.select. Areas communications 1997.

[10] Packet Scheduling Survey and Tutorial - Guansong Zhang

[11] Modeling Packet Scheduling Algorithms in IP Routers

[12] Xiao Qian, Mohamed Alghamdi, Mais Nijim, Ziliang Zong, Kiranmai Bellam, and Adam Manzanares, "Improving Security of Real-Time Wireless Networks Through Packet Scheduling" ,IEEE Transactions on Wireless Communications, Vol. 7, No. 9, September 2008.

[13] Xiaomin Zhu , Hao Guo, Shaoshuai Liang, Xiaoling Yang, An improved security-aware packet scheduling algorithm in real-time wireless networks, Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha, 410073, PR China

[14] Wireless Sensor Network, 2016, 8, 77-84 Published Online May 2016 in SciRes. http://www.scirp.org/journal/wsn http://dx.doi.org/10.4236/wsn.2016.85008.

[15] Xiaomim Zhu, HaoGuo, Shaoshuai Liang, Xioling Yang,"An improved Security-aware packet scheduling Algorithm", Information Systems Engineering Laboratory, National University of Defense, PR China.

[16] Arun Raj, P.Blessed prince, "Round Robin based secure aware packet scheduling in wireless networks, IJEST, ISSN:0975-5462, vol 5, No. 3, March 2013.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Nearest Keyword Set Search Queries on Multi-Dimensional Datasets

Mr. Aju P Benny,
M.Tech. Scholar,
Department of CSE,
Malla Reddy College of Engineering
Secunderabad, India

Dr. V. Bhoopathi,
Professor,
Department of CSE,
Malla Reddy College of Engineering
Secunderabad, India

Abstract – **Data mining is an important aspect in refining the data, this makes the data simpler and easy to use and it is a widely used technique to extract the useful information from huge chunks of data. The application developed is a user friendly application which helps to extract the informative keywords from text files and associates keywords with the file and the searching and retrieval of file is made easier with nearest keywords in a multidimensional data set of files. The integration of data mining concepts along with the cloud storage security techniques lead to the development of an efficient search application to retrieve files from the repository using easily remembered keywords.**

Keywords: Data security, keyword extraction, multidimensional data set.

I.INTRODUCTION

Present world is a technology centric world where each and every small works carried out in day to day life is dependent on the technology. Technology is rapidly growing such that it has occupied major portion in human routine. In this technological era the dependency of people over the information is growing hand in hand with the technology. Each and every activity is dependent on accessing the information and processing with it to get the required result. Starting from a small home to large organisation storing and retrieving of information has become an inseparable part of one's routine. Researches' have stated that on a whole the collection of data includes 10% of structured data and the remaining 90% are unstructured data. Information retrieval from structured data set is simpler compared to information retrieval from unstructured datasets. In order to discover the hidden patterns in the unstructured dataset data mining techniques are used.

Data mining is the process in which the patterns are extracted from the data set and the discovered patterns are analysed and used in studies. Machine learning, statistical analysis are some of the domains which uses the data mining techniques. In this application the data mining is applied over the text documents. The text documents with varied contents without any specific patterns to extract are being considered. A multidimensional data is the type of data in which there is heterogeneous data type objects that are grouped under certain attributes. But when the text document with no specified patterns are considered the pattern recognition becomes impossible, in such cases the Keywords extracted from every document is one dimension of the document that describes the feature of the document.

II. MOTIVATION

Over time the collection of documents increases and remembering all file names is impossible for normal human brains and retrieval of file is impossible if file name is unknown, but one can actually be aware of the searching content that is present in a file. In this application the searching based on the keywords document with the associated keywords are easier to search as data consumer can easily interpret about the content of the document rather than the file name. An application to search the documents along with the secured storage for documents is very necessary in every organisation and its implementation is shown in section four and the literature study is shown in section three, design in section four and conclusion in section six.

III. LITERATURE SURVEY

First technique is and it mainly on computing exact nearest and farthest neighbour which is a challenging task, especially in the case of high-dimensional data. Many techniques are used to solve the nearest neighbour problem but not much importance is on farthest neighbour problem. By the calculation of the farthest neighbour a clear idea is obtained for the elimination of unrelated objects to the query there by it helps in giving the result more accurately.[1]

Multidimensional text cube analysis is another technique which is used to analyse the textual documents and the analysis is done by applying the data mining technique over the documents in order to extract the hidden patterns out of it.[2]

Collective spatial keyword technique is another technique which is used to derive the result such that more than one object is required to satisfy the user's query in such cases one node is being considered as owner and other two nodes as sub objects that matches closely with query keyword. This is helpful in developing navigational search query applications where more than one object is necessary to satisfy the user's need.[3]

Cloud computing being one of the blooming technologies provides various services one such facility is the storage at minimum cost which makes most of the

organizations to store their data inside the cloud but there are chances of cloud being vulnerable to attacks. There is a technique called n-keyword search where the n is total number of distinct keywords present in all the documents and scalar product is performed over the co-ordinate keywords and homomorphic key exchange is done between the participating entities such that transmission is done securely. [4]

The literature survey of these papers helped in developing an idea to implement a search application over multidimensional data set.

IV.SYSTEM DESIGN

Any project before being implemented must be designed such that it gives a complete view of the entire project. The workflow is the main part of project design which is step wise explanation to the flow of project. The flow diagram for the project is shown below in Figure 1.

Figure 1 shows the work flow of admin module. There are six operations carried out at the time of uploading a file. The specified file is selected and file is being sent in to stop

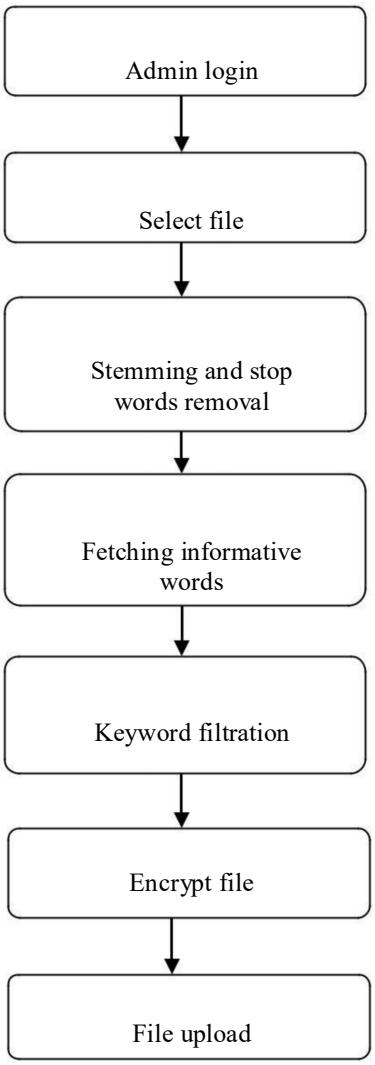


Figure 1 Work flow of admin

words removal and stemming process informative keywords are fetched and keyword filtration is done. The file is uploaded along with its associated keywords. The workflow of user is shown in figure 2 once user registers and gets the user ID and password and decryption key to the mail box from admin, then user can search the file by decrypting the key given to him and search file with the keywords and download if needed.

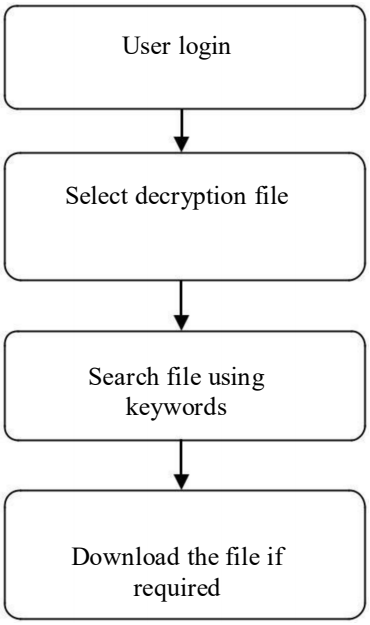


Figure 2 Work flow of user

Algorithm for upload process

Step 1: Start
Step 2: Read File (F) and access permission based on category(C).

Step 3: Remove unnecessary words and special characters.
Step 4 : Shortlist the Keywords.
Step5:UsingTermFrequency(TF) calculates weight age for Keywords.

Step 6: Let N be the number of category allowed to access.
Step 7: For I=1 To N.
Step 8: Fetch the hash key of Ith category.
Step 9: Using hashing technique with fetched hash key,

generate keyword hash tags for all the keywords.

Step 10: Insert all keyword hash tag into index.
Step 11: Repeat from Step 7 to Step 10 up to I=N.
Step 12: Stop

Algorithm for search process

Step 1: Start
Step 2: Get the Keyword (K) from User (U).
Step 3: Find the User Category (UC).
Step 4: Fetch Hash key for UC.
Step 5: Generate Trapdoor using UC hash key.
Step 6: Search trapdoor on index array.
Step7: Filter all the matched index elements.
Step 8: Shortlist filter from filtered index.
Step 9: Using Inverse document Frequency Rank the files.
Step 10: Display the file list to user.
Step 11: Stop

ISBN 978-81-939386-1-4

The above are the algorithms for upload and search processes respectively. The upload process is carried out by the data provider and downloading happens by the data consumer.

V. PROPOSED MODEL

The system architecture of the search application developed is shown in figure 3. The two main modules are admin and the user. This is a multiuser environment based application. Once the user is authorised the user can get access to the files through the internet if the application is installed in the user's system. Admin is the data owner who maintains the user details and uploads the files in to the database. Once the registration is done the user gets a decryption sent to the mail id. If a user wants to retrieve the files then the decryption key sent to the user's mail is decrypted then the search box opens up and the user can search the file using the predicted keywords.

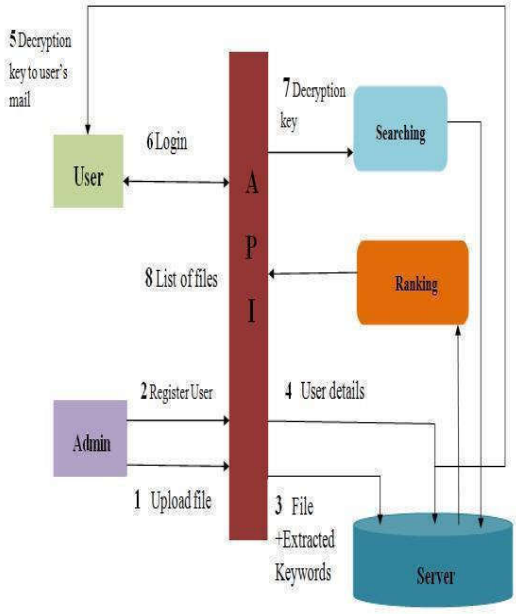


Figure 3: System architecture

Admin being the data owner maintains records and at the time of uploading files the admin is given the authority to select the grade such that the specific file being uploaded can be viewed only by the user who belongs to same grade. This facility is provided with a security point of view and also to reduce the searching process which happens only within the particular grade to which user belongs. The searching process need not consider all the files inserted in the database instead it is sufficient to carry out the searching only within the particular grade. Once the user logs in the user needs to decrypt the key which is provided by the admin at the time of the registration. Once the decryption of key is done then the search box is being opened to the user, where the user can type the keywords either multiple or single keyword. Once the matching of keywords is done the respected files are displayed such that the file with the highest rank will be at the top followed by files with least ranking.

VI. RESULT ANALYSIS

The result of the work carried out is to get the list of files based on the keywords weightage in the form of ranked list as shown in figure 4. The test cases for the work are shown in figure 5.

Ranked Files from Server

S.NO	File Name	Word %	Download
12	w.txt	90.0	Download
13	t.txt	30.0	Download
14	q.txt	30.0	Download
15	p.txt	30.0	Download

Figure 4 : List of files

Test cases	Input	Result
Keywords	Keywords matched with index keywords	PASS
	Keywords spelt wrong or not matched	FAIL
Number of keywords	Keywords less than five	PASS
	Keywords more than five	FAIL
Type of documents	Text file	PASS
	Pdf, word document, excel, image file.	FAIL

Figure 5 : Table of test cases

The analysis is done based on the uploading and downloading time. The uploading time take for any file size is always greater than the downloading time. Uploading takes more as the stemming and stop words removal are happening at the time of uploading. And the downloading time is lesser than the uploading time this because there is no much steps involved in downloading a file hence file search is efficient using this application.

The time recorded for various sizes is shown in figure 6 and the bar chart in figure 7 clearly depicts that for any size of files the uploading time is always greater than downloading time.

VII. CONCLUSION

The application developed is useful in any organization where the huge amount text data is stored in the form of files, so late at the time of retrieval the work of user becomes easier by searching the file with the keywords. The application can further be developed to take inputs such as images, audio, video and documents other than text files and also the application can be developed in to an android app which reduces the burden of the user to a greater level.

VIII. RRFERENCE

[1] Yifan Hao, Huiping Cao, Yan Qi, Chuan Hu, Sukumar Brahma, Jingyu HanNew Mexico State University, Las Cruces, "Efficient Keyword Search on Graphs using MapReduce" 2015 IEEE International Conference on Big Data (Big Data)

[2] Cheng Long Raymond Chi-Wing Wong Ke Wang Ada Wai-Chee Fu The Hong Kong University of Science and Technology Simon Fraser University The Chinese University of Hong Kong "Collective Spatial

[3] Shafin Rahman Department of Electrical & Computer Engineering North South University Dhaka, Bangladesh Mrigank Rochan Department of Computer Science University of Manitoba Winnipeg, Canada "Fast Farthest Neighbor Search Algorithm for Very High Dimensional Data" 19th International Conference on Computer and Information Technology, December 18-20, 2016, North South University, Dhaka, Bangladesh

[4] Fangbo Tao, Kin Hou Lei, Jiawei Han, ChengXiang Zhai, "EventCube: Multi-Dimensional Search and Mining of Structured and Text Data" KDD'13, August 11–14, 2013, Chicago, Illinois, USA. Copyright 2013 ACM 978-1-4503-2174-7/13/08

[5] Gandeevan Raghuraman Pooja Nilangekar Pallavi Vijay Kavya Premkumar Saswati Mukherjee "Cloud based Privacy Preserving EfficientDocument Storage and RetrievalFramework"1234Department of Computer Science,Anna University,Chennai 5Department of Information Technology, Anna University Chennai 978-1-4673-8200-7/15/ ©2015 IEEE

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

AchievingEfficientandSecureDataAcquisition for Cloud-supportedInternetofThingsinSmart Grid

SOUNDARYA GUDIPUDI,
M.Tech Scholar,
Dept of Computer Science & Engineering,
Malla Reddy Colelge of Engineering,
Hyderabad
nandugudipudi222@gmail.com

DR.V.BHOOPATHY,
PROFESSOR,
Dept of Computer Science & Engineering,
Malla Reddy College of Engineering,
Hyderabad
v.bhoopathy@gmail.com

Abstract— **Cloud-supported Internet of Things (Cloud-IoT) has been broadly deployed in smart grid systems. The IoT front-ends are responsible for data acquisition and status supervision, while the substantial amount of data is stored and managed in the cloud server. Achieving data security and system efficiency in the data acquisition and transmission process are of great significance and challenging, because the power grid-related data is sensitive and in huge amount. In this paper, we present an efficient and secure data acquisition scheme based on CP-ABE (Ciphertext Policy Attribute Based Encryption). Data acquired from the terminals will be partitioned into blocks and encrypted with its corresponding access sub-tree in sequence, thereby the data encryption and data transmission can be processed in parallel. Furthermore, we protect the information about the access tree with threshold secret sharing method, which can preserve the data privacy and integrity from users with the unauthorized sets of attributes. The formal analysis demonstrates that the proposed scheme can fulfill the security requirements of the Cloud-supported IoT in smart grid. The numerical analysis and experimental results indicate that our scheme can effectively reduce the time cost compared with other popular approaches.**

I. INTRODUCTION

with the support of modern information technologies like the Internet of Things (IoT) and cloud computing, smart gridhasemergedasthenext-generationpowersupplynetwork, in which the electricity is generated according to the real-time demands of electric equipment or household appliances [1,2]. Tomakethesmartgridmoreintelligent,agreatnumberofIoT terminals are deployed to gather the status of the power grid timely for the control center. Some sample applications are such as the power transmission line monitoring, power generation monitoring, substation state monitoring, smart metering, electric energy data acquisition, smart home. For instance, in power transmission line monitoring scenario, using preplaced sensors, the status parameters of the transmission line and power towers can be gathered in real time, so that any fault can be diagnosed and locatedinatimelymanner. In smart grid, the different kinds of applications mentioned above all generate an enormous amount of data, which needs tobestoredandmanagedefficiently.Cloud-IoTisproposedto addressthisissue[2,3].AsshowninFig.1,withthesupportof cloud computing, mass data from different IoT terminals can be collected and processed by local front-end servers, then transferred and stored in the cloud servers. The data in cloud canbeaccessedbyvarioustypesofdatausers.Thepowergrid staff can continually monitor the status of power grid. Researchers andgovernment agencies cananalyzethe datafor researchorpolicymaking.

Fig.1IllustrationofcloudsupportedIoTinsmart grid Actually, some works on Wireless Sensor Networks can be used for reference, such as [4-9]. However, there are still severalproblemsandchallengesinsmartgriddataacquisition. First, the efficiency of data acquisition should be considered due to the large amount of data to be encrypted/decrypted and transferred. It's critical to ensure an acceptable the data acquisition time. Second, the protection of data security and privacymustbekeptinmind.Todealwiththesetwoproblems

simultaneously, in this paper, we present an efficient and secure data acquisition scheme based on CP-ABE. The main contributionsofourworkcanbesummarizedasthefollowing:

EXISTING SYSTEM

With the support of modern information technologies like the Internet of Things (IoT) and cloud computing, smart grid has emerged as the next-generation power supply network, in which the electricity is generated according to the real-time demands of electric equipment or household appliances.To make the smart grid more intelligent, a great number of IoT terminals are deployed to gather the status of the power grid timely for the control center. Some sample applications are such as the power transmission line

monitoring, power generation monitoring, substation state monitoring, smart metering, electric energy data acquisition, smart home. For instance, in power transmission line monitoring scenario, using preplaced sensors, the status parameters of the transmission line and power towers can be gathered in real time, so that any fault can be diagnosed and located in a timely manner. In smart grid, the different kinds of applications mentioned above all generate an enormous amount of data, which needs to be stored and managed efficiently. Cloud-IoT is proposed to address this issue .with the support of cloud computing, mass data from different IoT terminals can be collected and processed by local front end servers, then transferred and stored in the cloud servers. The data in cloud can be accessed by various types of data users. The power grid staff can continually monitor the status of power grid. Researchers and government agencies can analyze the data for research or policy making.

PROPOSED SYSTEM

In this paper, we present an efficient and secure data acquisition scheme based on CP-ABE (Ciphertext Policy Attribute Based Encryption). Data acquired from the terminals will be partitioned into blocks and encrypted with its corresponding access sub-tree in sequence, thereby the data encryption and data transmission can be processed in parallel. Furthermore, we protect the information about the access tree with threshold secret sharing method, which can preserve the data privacy and integrity from users with the unauthorized sets of attributes. The formal analysis demonstrates that the proposed scheme can fulfill the security requirements of the Cloud-supported IoT in smart grid. The numerical analysis and experimental results indicate that our scheme can effectively reduce the time cost compared with other popular approaches.

IMPLEMENTATION

MODULES:

- Data Owners
- Data Requester/Receivers
- Cloud Servers
- Attribute Authority

MODULES DESCRIPTION:

Data Owners: In this module, the Data Owner decides the access policy and encrypts the data with CP-ABE. The encrypted data will be uploaded to the Cloud Servers. The data owner response the data receivers file requests. Data Owner is assumed to be honest in the system.

Data Requester/Receivers: In this module, the Data Receivers send the decryption request to data owner and obtain the ciphertexts over the internet. Only when their attributes satisfy the access policies of the ciphertext, can they get access to the plaintexts. Data requester/receivers may collude to access the data that is otherwise not accessible individually.

INPUT DESIGN AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

VI. CONCLUSION

Cloud-IoT techniques are widely deployed in Smart Grid. Huge amount ofdatais gathered byIoTfront-enddevices and stored inthe back-end cloud servers. However, achieving data security and system efficiency in the data acquisition and transmissionprocessareofgreatsignificanceandchallenging.
Existing related schemes cannot deal with this challenging issue well. To tackle with this problem, we propose a secure and efficient data acquisition scheme for Cloud-IoT in smart grid. Inthe proposed scheme, the large data is partitioned into several blocks, and the blocks are encrypted/decrypted and transmitted in sequence. In addition, we adopt the dual secret sharing scheme, which realizes the privacy-preserving, the data integrity check and the attributes check simultaneously. The analysis shows that the proposed scheme can meet the security requirements of data acquisition in smart grid, and it Also reduces response time overhead significantly compared to other popular s charms.The data of the proposed scheme is not uploaded in real time, it is offline before encryption. The researchondatatimeliness will be our future work. .

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid—The new and improved power grid: A survey", *IEEE communications surveys & tutorials*, vol. 14, no. 4, pp. 944-980, Dec. 2012.

[2] J. Singh, T. Pasquier, J. Bacon, and H. Ko, "Twenty Security Considerations for Cloud-Supported Internet of Things", *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 269-284, Dec. 2016.

[3] L. Jiang, L. D. Xu, H. Cai, Z. Jiang. "An IoT-Oriented Data Storage Framework in Cloud Computing Platform " , *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1443-1451, Feb. 2014.

[4] X. Du, Y. Xiao, M. Guizani, and H. H. Chen, "An effective key management scheme for heterogeneous sensor networks", *Ad Hoc Networks*, vol. 5, no. 1, pp. 24-34, Jan. 2007.

[5] X. Du, and H. H. Chen, "Security in wireless sensor networks ", *IEEE Wireless Communications*, vol. 15, no. 4, pp. 60-66, Aug. 2008.

[6] X. Du, M.Guizani, Y. Xiao, and H. H. Chen, "Defending DoS Attacks on Broadcast Authentication in Wireless Sensor Networks ", *in proc. IEEE ICC*, 2008, pp. 1653-1657.

[7] X. Hei, X. Du, J. Wu, and F. Hu, " Defending Resource Depletion Attacks on Implantable Medical Devices ", *in proc. IEEE GLOBECOM*, 2010, pp. 1-5.

[8] X. Hei, and X. Du, "Biometric-based two-level secure access control for Implantable Medical Devices during emergencies", *in proc. IEEE INFOCOM*, 2011, pp. 346-350.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Privacy and Integrity Preserving Top-*k* Query Processing for Two-Tiered Sensor Networks

KARROLLA BHARATH
PG STUDENT,
Department of Computer Science & Engineering,
Malla Reddy College of Engineering
Hyderabad, India
karrollabharath22@gmail.com

M. NARAYANAN
PROFESSOR
Department of Computer Science & Engineering,
Malla Reddy College of Engineering
Hyderabad, India
narayanan.cse@mrce.in

Abstract— **Privacy and integrity have been the main road block to the applications of two-tiered sensor networks. The storage nodes, which act as a middle tier between the sensors and the sink, could be compromised and allow attackers to learn sensitive data and manipulate query results. Prior schemes on secure query processing are weak, because they reveal non-negligible information, and therefore, attackers can statistically estimate the data values using domain knowledge and the history of query results. In this paper, we propose the first top-*k* query processing scheme that protects the privacy of sensor data and the integrity of query results. To preserve privacy, we build an index for each sensor collected data item using pseudo-random hash function and Bloom filters and transform top-*k* queries into top-range queries. To preserve integrity, we propose a data partition algorithm to partition each data item into an interval and attach the partition information with the data. The attached information ensures that the sink can verify the integrity of query results. We formally prove that our scheme is secure under IND-CKA security model. Our experimental results on real-life data show that our approach is accurate and practical for large network sizes.**

***Index Terms*— Two-tiered sensor networks, privacy preserving, top-k queries.**

I. INTRODUCTION

A. Motivation

TWO-TIERED sensor networks have been widely adopted for their scalability and energy efficiency. A large number of sensors [1]-[5], equipped with limited storage and computing capacity, are deployed in fields. Some storage nodes, equipped with large storage and powerful computating capacity, are deployed among sensors for storing measurement data from the neighboring sensors, as shown in Figure 1. A sink serves as a terminal device that sends queries to the storage nodes and retrieves the sensor data of interest. Due to the importance of two-tiered sensor network architecture, several commercial storage nodes, such as StarGate [6] and RISE [7], have also been developed.

The storage nodes offer two major benefits compared to an unstructured sensor network model. First, the storage nodes are responsible for the collection, storage and transmission of the sensory data from the sensors to the sink. The sensors save a significant amount of energy by eliminating sensor to sensor relay transmissions towards the sink and prolong the life of the network. Second, the storage nodes have more computing power and storage capacity than the sensors. Therefore, the sink can issue complex queries, such as the range or top-*k* queries, to retrieve several data items in a single query. This saves the sensor nodes' energy and network bandwidth required for answering the sink queries. However, due to their importance in network operations, the storage nodes are more vulnerable to attack and compromise. Attackers can not only steal the sensitive information on the storage node, but also leverage the query processing functionality of the storage node to feed false information to the sink.

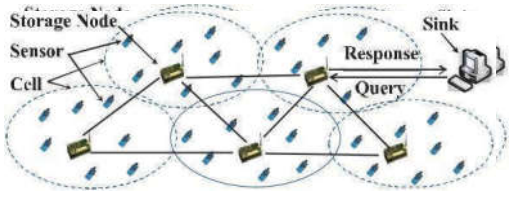


Fig. 1. Architecture of two-tired sensor networks.

B. Problem Statement

We address the problem of privacy and integrity preserving top-*k* queries in two-tiered sensor networks to protect against storage node compromise. Our goal is to design scheme to enable storage nodes to process top-k queries correctly without knowing the actual value of

data stored in them and allow the sink to detect misbehavior of storage nodes. Top-k query processing, *i.e.*, finding the k smallest or largest data items col-lected from a specified sensed area, is a fundamental operation in sensor networks [8]–[10]. Such top-k queries enable users to get their most desired environmental information such as pollution index, temperature, humidity and so on. Our choice
of the top-k query problem is motivated from the fact that it can be viewed as a generalized version of range query, which allows the sink to learn several values with a single query. We consider a storage node N_i and a set of sensors in
its neighborhood $\{S_1, S_2, \cdots, S_g\}$ where each S_j collects a set of data items $D(S_j) = \{d^j_1, d^j_2, \cdots, d^j_c\}$. When the sink issues a top-k query to N_i, the storage node responds with
the top-k largest or smallest data items from the stored data: $\cup^{g}_{j=1} D(S_j)$. Now, to preserve privacy of data in the event of storage node compromise, a sensor S_j encrypts the each sensed data item d^j_x into $(d^j_x)_{k_j}$ using a secret key k_j shared with the sink and builds an index I_j for $D(S_j)$. Then, S_j
submits the index I_j along with the set of encrypted data items $\{(d^j_1)_{k_j}, (d^j_2)_{k_j}, \cdots, (d^j_c)_{k_j}\}$ to the storage node N_i. With this background, we state the problem of privacy and integrity
preserving top-k queries as follows.

For a top-k query, the sink generates a trapdoor t_k for k, which is the query condition in an encrypted form, and sends t_k to N_i. Based on t_k and the indexes $I_1, I_2, \cdots, I_a$, N_i should be able to determine a set of encrypted data items satisfying the query condition without having to decrypt either t_k or
$\{(d^j_1)_{k_j}, (d^j_2)_{k_j}, \cdots, (d^j_c)_{k_j}\}$. To preserve integrity, given the top-k results from a storage node N_i, the sink should be able
to verify the validity of the results using some information embedded in the top-k query results.

C. Adversary and Security Model

In this work, we adopt the *semantic security against chosen-keyword attack*, *i.e.*, *IND-CKA* security model pro-posed in [11], which has been widely accepted in privacy preserving techniques like [12]–[15]. Specifically, our model uses the refined simulation based definition by Curtmola *et al.* in [14]. Using the standard definitions from [16], we consider
a probabilistic polynomial-time (PPT) adversary who attempts to leak the privacy of the secure index I of a given set of data items. The adversary is allowed to generate a sample *history*, $H = \{D, d\}$ where D is a set of data items of the adversary's choice and d is a polynomial sized set of chosen queries. Now, if there exists a polynomial-time simulator that can simulate the actions of a secure index given only H, then the secure index is *semantically secure against adaptive chosen keyword attacks*. We consider the case where the adversary is not allowed to see the secure index or any secure trapdoors
or any data items, prior to generating H and this adversary is called *non-adaptive* adversary [14].

Briefly, the simulation proceeds thus: the adversary submits H to a challenger, who has access to a secure index algorithm and a simulator, and either generates a secure index using D or simulates the secure index using the simulator. In the challenge phase, the adversary obtains trapdoors for d from either the real index or the simulator and is asked to distinguish between them with non-negligible probability [16]. The existence of such a simulator proves the security of the index scheme and we show such a simulator for our index in this work.

D. Limitations of Prior Art

Previous works that are in two-tiered sensor networks and closely related to ours are privacy preserving top-k
queries [17], [18] and privacy preserving range queries [3], [19]–[22]. Broadly, these works fall into two categories: buck-eting schemes [3], [20], [22] and order preserving encryp-tion based schemes [17]–[19], [21]. Note that we classify SafeQ [19] into order preserving scheme because in SafeQ scheme, sensors sort their measurement data before encryption and the encrypted data items are kept in an ordered manner in the storage nodes to facilitate the secure queries.

The query result includes all data items that their corresponding bucket IDs are in the trapdoor. However, the privacy protection of bucketing schemes is weak because attackers can statistically estimate the actual value of both the data items and the queries using domain knowledge and historical query results, as pointed out in [23]. Also, the communication cost is high in these schemes because a query result includes many false positive data items.

Order preserving encryption schemes describe techniques to preserve the relative ordering of data even after data encryption. Given any two data items d_1 and d_2, and an order preserving encryption function f, $f(d_1) \leq f(d_2)$ if and only if $d_1 \leq d_2$. For each $d^j_i \in D(S_j)$, sensor S_j computes $f(d^j_i)$ and submits it with $(d^j_i)_{k_j}$ to a storage node. Note that different sensors may have different order preserving encryption functions [18], [21]. In [24], the authors pointed out that the privacy preserving schemes based on order preserving encryptions are not secure.

E. Technical Challenges and Proposed Approach

Our approach for privacy and integrity preserving top-k query processing consists of a novel approach to sensor data processing that we combine with symmetric key-based techniques to achieve the desired data privacy and integrity verification. Our data processing approach helps to solve the two important challenges of privacy preserving *top-k querying* over arbitrary distributions of sensor data and *integrity verification* of the query results.
. To address this challenge, we note that a top-k query can be approximated by an appropriate range query and the uniformly distribution of sensor data facilitates this approximation. First, we describe a data distribution transformation method to transform the arbitrary sensor data distribution into an approx-imate uniform distribution. Second, to preserve privacy of sensor data, we describe an algorithm to build an IND-CKA

secure privacy preserving index on the transformed data using pseudo-random one-way hash functions and Bloom filters. Third, using prefix based techniques, we transform *lesser than* and *greater than* comparisons into *equality* checking, which only involves set membership verification operation on the Bloom filter indexes.

Finally, we describe a range estimation algorithm to transform a top-k query into a special range query, called *top-range* query, and enable the storage node to process the query on the secure Bloom filter indexes.

The second challenge is to verify the integrity of the query results. Towards this, we propose a novel data partition algorithm to partition the data items into intervals. We describe an embedding technique to embed the interval information into the corresponding data items before encrypting them.

Our proposed scheme addresses the limitation of prior arts by providing IND-CKA security through a combination of range approximation and data index generation techniques. The index indistinguishability prevents attackers to statistically infer data values from the query processing information.

F. Key Contributions

The key contributions of our work are four fold. First, we propose the first privacy and integrity preserving top-k query processing scheme in two-tiered sensor networks that is secure under IND-CKA security model. Second, we describe a novel approach to sensor data transformation and partitioning to index the data appropriately. Third, we propose a query transformation algorithm to transform top-k queries into top-range queries. Fourth, we implemented and evaluated our scheme thoroughly on a real data set. Experimental results show that our scheme is efficient and scalable.

II. Related Work

Related schemes to our approach can be found similar in cloud computing and database domains. These works can be divided into three classes: bucketing schemes, order preserving schemes, and public-key schemes. Hacigumus *et al.* proposed the first bucket partition scheme [25] to query encrypted data items without allowing the server to know the exact data values. Hore *et al.* investigated the problem of optimal bucket partitioning and proposed two secure query schemes, one for one-dimensional data [23] and the other for multi-dimensional data [26]. Agrawal *et al.* adopted the bucketing scheme's idea and proposed an order preserving scheme to further protect data privacy [27]. Boldyreva *et al.* proposed two order preserving schemes [28], [29]. However, these schemes have the security weaknesses as discussed in Section I-D. In [30], Li *et al.*, propose a privacy preserving range query processing scheme for outsourced data items in cloud computing, which is proved to be secure under IND-CKA security model. However, this scheme cannot perform top-k querying with intergrity verification for sensor networks.

III. System Model and Assumptions

We adopt the system model used in existing privacy preserv-ing querying approaches [3], [20]–[22] for two-tiered sensor networks.

First, we assume that, all the sensors and the storage nodes are loosely synchronized. Under this assumption, we divide the time into a series of fixed length time slots. In each time slot, a sensor collects multiple integer data items, whose minimal and maximal possible values are known. At the end of a time slot, the sensor submits these data items to its closest storage node. Second, each sensor shares two symmetric keys with the sink: a common key and a secret key. The common key is shared among all the sensors and the sink. The secret key is shared between a given sensor and the sink. These two keys are stored in tamper-proof hardware and therefore, would not be compromised even if the sensors are captured by attackers. Third, the sensors may collect multi-dimensional data. In this case, the sensors compute a score for each multi-dimensional data, and then allow the sink to query on these scores, an approach adopted in previous approaches [10], [17], [18]. Without loss of generality, we assume that top-k queries are performed on one-dimensional data and the sink is interested
in the k smallest data items in the sensed data.

Let $d_1^j, d_2^j, \cdots, d_n^j$ denote the data items collected by

sensor S_j in a time slot. For all sensors $S_1, S_2, \cdots, S_g$ that are close to a storage node N_i, we have:

$$\sum_{j=1}^{g} n_j = n, \quad \text{(III.1)}$$

which means totally n data items are submitted to N_i after the time slot. We define a lower bound variable d_0 and an upper bound variable d_{n+1}, such that for all j, $d_0 \leq \min\{d_1^j, d_2^j, \cdots, d_{n_j}^j\}$ and $\max\{d_1^j, d_2^j, \cdots, d_{n_j}^j\} \leq d_{n+1}$. The possible range for all data items is $[d_0, d_{n+1}]$.

We represent a data item d as a w-bit binary string $b_1 b_2 \cdots b_w$ where b_i $\forall i \leq w$ denotes the bit-value at position i in the string. A k-prefix of d is the first k bits of d and the remaining $w - k$ bits are not cared, *i.e.*, $*$s, which can be 0 or 1. For example, for the numeric value 4 represented in 4-bits as 0100, the 2-prefix is denoted by 01 $**$. A prefix $\{0, 1\}^k\{*\}^{w-k}$ denotes the interval $[\{0, 1\}^k\{0\}^{w-k}, \{0, 1\}^k\{1\}^{w-k}]$. For example, 01 $**$ denotes [0100, 0111], *i.e.*, the interval between the minimum and the maximum values that this prefix denote. The prefix family of d, denoted by $F(d)$, is defined as the prefix set generated by replacing each right-most bit value by $*$; *i.e.*, $F(d) = \{b_1 b_2 \cdots b_w, b_1 b_2 \cdots b_{w-1}*, \cdots, b_1 * \cdots *, ** \ldots *\}$. For example, for the value 4 represented in 4-bits, the prefix family is: $F(4) = \{0100, 010*, 01 **, 0 ***, *** *\}$. For any k-prefix, $b_1 b_2 \cdots b_k * \cdots *$, we define two prefixes, *child-0* prefix, obtained by replacing first $*$ by 0, *i.e.*, $b_1 b_2 \cdots b_k 0 * \cdots *$ and *child-1* prefix, obtained by replacing first $*$ by 1, *i.e.*, $b_1 b_2 \cdots b_k 1 * \cdots *$. For example, for 01 $**$, the child-0 prefix is 010$*$ and the child-1 prefix is 011$*$. We use p^0 and p^1 to denote child-0 and child-1 prefixes of p, respectively.

Finally, we define a *head prefix* as a prefix with a 0 before the first $*$; e.g., 010$*$ and a *tail prefix* as a prefix with a 1 before the first $*$; e.g., 011$*$

IV. Sensor Data Pre-Processing: Mapping and Partitioning

Our data processing involves four important steps: approx-imating uniform data distribution, data partitioning, interval embedding and index selection. First, we describe a method to transform the arbitrary sensor data into an approximate uniform data distribution. Second, we describe a data parti-tioning algorithm, which enables us to generate the neces-sary information for integrity verification. Third, we show the method to embed the interval information required for integrity verification of the query results. At last, we show the index value selection for building the secure index for top-k querying.

A. Approximating Uniform Distribution

Our approach to executing top-k queries is to view the sensor data into a uniformly distributed data and execute specially crafted range queries, which we call *top-range* queries. As sensor data follows an arbitrary distribution, it is difficult to execute top-k queries without the use of expensive homomorphic operations or order preserving encryption that enable comparison on the encrypted data.The major advantage of this transformation is that the sink node can transform a top-k query by an appropriately chosen range query, a top-*range* query, which will retrieve the expected top-k data items.

For achieving the desired transformation, we adopt the well-known approach from [27] while ensuring that it suits the sensor network domain. To approximate the arbitrary sensor

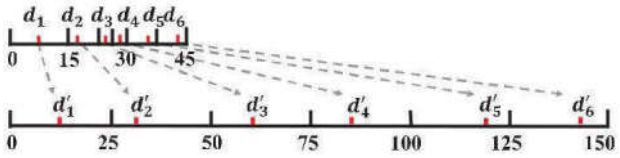


Fig. 2. Approximation to uniform distribution.

data distribution, say over $[d_0, d_{n+1}]$, to a uniform distribution, say over $[d_0, d_{n+1}]$, the approach in [27] is to define a uniform distribution over $[d_0, d_{n+1}]$ with the same number of data items in $[d_0, d_{n+1}]$ and map the new data items to the arbitrarily distributed sensor data items in $[d_0, d_{n+1}]$. First, we partition $[d_0, d_{n+1}]$ into multiple intervals such that every interval contains the same number of data items. Different intervals could vary in length. Second, for a given $[d_0, d_{n+1}]$, we partition this range into the same number of intervals as we partition $[d_0, d_{n+1}]$. In this partition, every interval has the same length. Third, we project the data values in $[d_0, d_{n+1}]$ to new data values in $[d_0, d_{n+1}]$, according to a projection relationship [27] between the corresponding intervals.

Figure 2 illustrates the idea of data distribution transformation. After the transformation, the n new data

values in $[d_0 , d_{n+1}]$ form an approximate uniform distribution. We note down the trans-forming relationship as a table and store it in each sensor. Thus, a top-k query can be approximated by a range query that fetches roughly k data items.

Note that, using this transformation as part of our privacy preserving scheme does not make our final scheme vulnerable to the privacy concerns inherent in [27], due to the following reasons. First, each sensor constructs a privacy preserving index (cf. Section V) using these new data values. As our indexes are IND-CKA secure, an adversary cannot reveal the contents of the index with non-negligible probability. Also, the sink and sensors can agree upon the transformation at network initialization or can exchange this information securely using the common network secret key.

B. Data Partitioning for Integrity Verification

In this section, we describe our approach to verify the integrity of the results returned due to the top-k query by the sink node. Intuitively, our integrity verification approach is to partition the data into distinct intervals and securely embed the interval information along with the data item. Note that, due to the transformation of the sensor data into uniform distribution, the indexing is performed on the new index values derived from the uniform distribution. However, even though the top-k query will be executed on the new data items, the storage node will send back the corresponding original data items to the sink node. This implies that the sink node needs to be able to verify that received data items correspond to the mapped data items, on which the top-k query was executed. Since the sink node is aware of the uniform distribution mapping approach, it will be able to derive the corresponding uniform data values and subsequently, the corresponding interval information, which enables the verification of the embedded interval information in the received data items. We next give the details of partitioning.

Our data partition algorithm consists of two phases: *partition phase* and *merging phase*. In the *partition phase*, starting with the n distinct data items distributed in the range $[d_0 , d_{n+1}]$, we partition the range $[d_0 , d_{n+1}]$ into n intervals that satisfy the following conditions. (1) Each interval contains exactly one data item. (2) The union of the intervals is the range. (3) There is no overlap between intervals. In the *merging phase*, we merge the empty intervals with their neighboring intervals.

Partition Phase: The first phase of our partition algorithm is as follows. Suppose the data set to be partitioned is $S = \{d_1, d_2 , \cdots , d_n\}$. If $|S| > 1$, we first generate a prefix $p = \{*\}^w$. Then, we partition S into S_0 and S_1, where $S_0 = \{\forall d_i|(p^0 \in F(d_i)) \cap (d_i \in S)\}$ and $S_1 = \{\forall d_i |(p^1 \in F(d_i)) \cap (d_i \in S)\}$. If $|S_i| > 1$, we generate child-0 and child-1 prefixes for p^i and the partition process iteratively goes on, here $i = 0, 1$. Figure 3 demonstrates an example of the partitioning process. In this example, there are 6 4-bit data items $S = \{2, 5, 6, 7, 10, 11\}$. First, we generate prefix $p = ****$. Second, we compute $p^0 = 0***$ and $p^1 = 1***$ for p. Finally, we divide these 6 data items into $S_0 = \{2, 5, 6, 7\}$ and $S_1 = \{10, 11\}$ according to p^0 and p^1. Because both S_0 and S_1 contain more than one data item, the partition process continues.

At the end of this phase, we have 8 intervals as shown in Figure 3. And each interval is represented by a prefix.

Merging Phase: In the merging phase, we eliminate empty intervals by merging. If the prefix representing the empty interval is a head prefix, we merge this interval with its right-closest interval; if the prefix representing the empty interval is a tail prefix, we merge this interval with its left-closest interval. For example, in Figure 3, there are two empty intervals $E1$ and $E2$ represented by $11**$ and $100*$, respectively. The interval represented by $11**$ will be merged with the interval repre-sented by 1011; and the interval represented by $100*$ will be merged with the interval represented by 1010. Figure 4 shows the partition result of the data set $S = \{2, 5, 6, 7, 10, 11\}$

.

C. Embedding Intervals With Data

In this step, we embed the interval information, to which a transformed data item belongs, into the original data item sent to the sink. Suppose a data item d_j is transformed into d_j and partitioned into the interval R_j . The properties of prefix ensure that there are only two cases: (1) R_j is represented by one prefix; (2) R_j is represented by two prefixes. Considering these two cases, we use the form $d_j |(b_j , e_j)$ to represent the interval information, where b_j means the last b_j bits of d_j should be replaced by 0 to get the beginning value of R_j and e_j means the last e_j bits of d_j should be replaced by 1 to get the ending value of R_j . If d_j is a w-bit number, b_j and e_j only need log w+1 bits to represent. We use $N(p)$ to denote the number of $*$ in p. If R_j is represented by one prefix p, we let $b_j = e_j = N(p)$. If R_j is represented by two prefixes p_h and p_t and d_j is in the interval denoted by p_h, we let $b_j = N(p_h)$ and $e_j = N(p_t)+1$; otherwise, we let $b_j =$

$N(p_h)+1$ and $e_j = N(p_t)$. For example, in Figure 4, data item 2, whose
binary form is 0010, is in the interval represented by 00 ∗∗; we use 0010|(010, 010) to denote the interval; data item 11,
whose binary form is 1011, is in the interval represented by 1011, 11 ∗∗and 11 is in the interval denoted by prefix 1011, we use 1011|(000, 011) to represent the interval. The sensor S_i encrypts d_j with $d_j|(b_j, e_j)$ into the form $(d_j|d_j|(b_j, e_j))_{k_i}$. We use $E_i(d_j)$ to denoted $(d_j|d_j|(b_j, e_j))_{k_i}$ in the following
description. The sink decrypts $E_i(d_j)$ to recover the data value d_i and get the interval information via $d_j|(b_j, e_j)$. Next, we describe our data indexing strategy, which will allow the sink node to generate the necessary interval information for integrity verification.

D. Index Selection

In the following discussion, the data items under consid-eration are from the approximate uniform distribution, which resulted from the original sensor data items. To guarantee that the interval information needed for verifying the integrity of a query result is included in the result, we need to include the interval information of all data items covered by the top-
k query. Our approach is to select the lower bound value of the interval containing the index data item. Consider Figure 4 where six data items 2, 5, 6, 7, 10, 11 are partitioned into six intervals [0, 3], [4, 5], [6, 6], [7, 7], [8, 10], [11, 15], respectively, using our

partitioning algorithm. For instance, to index 2, we consider the interval [0, 3] and choose 0 as the indexing point. This is because the prefix 00 ∗∗, generated from the 4-bit binary representation of 0, can represent the entire interval [0, 3]. We use $d^m{}_i$ to denote the index of data d_i.

V. Privacy Preserving Index Generation

In this section, we describe our solution for generating the privacy preserving index upon which the sink can issue top-k queries without revealing the query contents. We present our solution using Bloom filters for indexing the data and keyed pseudo-random hash functions for preserving privacy.

A. Prefix Encoding and Bloom Filter Indexing

Note that, to execute the top-k query, our approach is to convert the top-k query in a specially crafted range query, *i.e.* top-range query. But, if each data item is represented by a numeric value, to execute a top-range query, we need three comparison operations, *less-than*, *bigger-than*, and *equal-to*, to test if this data item falls in the query range. To solve this problem, we adopt the prefix membership testing scheme in [33] in combination with the Bloom filter [34], which is a popular data structure for solving the prefix membership problem. The key advantage of the prefix membership testing approach is that we only need to check if a given query prefix is present inside a set of stored prefixes and therefore, we only need to implement the *equal-to* operation.

For a given range $[a, b]$, we can find a prefix set S such that the union of the intervals represented by the prefixes in S is equal to $[a, b]$. The *minimum prefix set* for $[a, b]$, $S([a, b])$, is defined as the prefix set of minimum cardinality such that the union of the prefixes is equal to $[a, b]$. For example,
$S([0, 4]) = \{00**, 1000\}$.

Following the definitions of *prefix family* and *minimum prefix set*, the following two properties hold. First, given a prefix p, an integer d falls into the interval represented by p if and only if $p \in F(d)$. Second, for a data item d and a range $[a, b]$, $d \in [a, b]$ if and only if there exists a prefix p, $p \in F(d)$ such that $p \in S[a, b]$. Therefore, for any integer data d and a range $[a, b]$, we can test whether $d \in [a, b]$ by testing whether

$$F(x) \cap S([a, b]) = \varphi. \quad \text{(V.1)}$$

For example, $3 \in [0, 4]$ and $F(3) \cap S([0, 4]) = \{00 ∗ ∗\}$. Based on this discussion, we now describe how the prefix membership testing is implemented using the Bloom filter structure [34]. A Bloom filter structure consists of an M-bit array, where the bits are initially set to 0 and a set of r independent hash functions h_i $1 \le i \le r$ such that, h_i :
$N \times N \rightarrow \{0, M - 1\}$.

For each data item d_j $(0 \le j \le n)$, we use a Bloom filter B_j to store the prefix family of its corresponding indexing value $d^m{}_j$, *i.e.*, we store $F(d^m{}_j)$ in B_j. Each $p \in F(d^m{}_j)$ is hashed with each of the r hash functions and the corresponding locations in the Bloom filter B_j are set to 1. To verify if a Bloom filter contains a queried data item, a verifier hashes the item using each of the r hash functions and tests if all the hashed locations in the Bloom filter are set to 1 or not.

But, directly storing the data items into Bloom filters can enable the storage node to breach the privacy of the stored data in two different ways. First, since the sensor data are bounded within a finite domain and given that the Bloom filter hash functions are public, the storage node can brute force the stored data items. Second, the storage node can compare the common bit locations of Bloom filters of different sensors and find out the common readings across the two Bloom filters.

Therefore, to prevent the privacy leakage due to these attacks, we describe a secure randomization approach that generates our privacy preserving index structure.

B. Randomizing Bloom Filter Indexes

In this phase, we describe an approach to store the index data item in such a way that it is not possible for a storage node to gain any information regarding the stored values using the brute-force or inference attacks described in the previous section. Note that, in time slot t, we assume that all sensors share an l-bit symmetric secret key k_t with the sink.

First, to protect against brute-force attack, we encode a secure value of a given prefix into the Bloom filter. For each prefix $p_i \in F(d_j)$, we compute a secure one-way hash, using the HMAC keyed-hash function and the secret key k_t, as follows $h_t(p_i) = \text{HMAC}_{k_t}(p_i)$, here h_t denotes the use of HMAC with k_t. The purpose of this step is to achieve one-way-ness, *i.e.*, given prefix p_i and the secret key k_t, it is computationally efficient to compute the hash result; but giving the hash result, it is computationally infeasible to reverse compute the secret key k_t and p_i. Now, instead of encoding the original prefix into the Bloom filter, we encode the securely hashed version. This prevents the brute-force attack as it is easy to compute $h_t(p_i)$ from p_i and h_t, but the vice-versa is not possible.

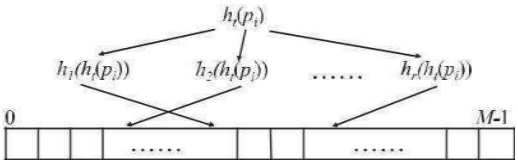
Second, to prevent inference attacks across two given Bloom filters, we devise an approach to encode a common prefix into different random locations across any two given Bloom filters. We generate r distinct numbers from the encryption of the corresponding data item. For a data item d_j, sensor S_i first computes the r hash values $\text{SHA}(1||E_i(d_j)), \cdots, \text{SHA}(r||E_i(d_j))$, where $||$ denotes catenation operation. Then, sensor S_i chooses the first l bits of each hash $\text{SHA}(k||E_i(d_j)$ to generate a unique number R_x, where $1 \le x \le r$. Assume that there are no two collected data items in a time slot are equal (the encryptions of two equal data items can be different by concatenating a measurement datum with its sequence number before encryption). Using these r numbers and HMAC, we construct r pseudo-random hash functions $h_1, h_2, \cdots, h_r$

where $h_x(.) = \text{HMAC}_{R_x}(.) \bmod M$ and M is the number of

cells in a bloom filter. We compute r hashes: $h_1(h_t(p_i))$, $h_2(h_t(p_i))$, $\cdots$, $h_r(h_t(p_i))$. Let B_j be an M bits bloom filter. For each prefix p_i and for each $1 \le x \le r$, we let $B_j[h_x(h_t(p_i))] := 1$. The purpose of choosing r unique

numbers is to eliminate the correlation among different Bloom filters. For the same prefix p, these unique random numbers allow us to hash p independently for different Bloom filters.

Without the use of these numbers, if prefix p_i is shared by $F(d_1)$ and $F(d_2)$ of two different data items d_1 and d_2, then for all the r locations $h_1(h_t(p_i))$, $h_2(h_t(p_i))$, $\cdots$, $h_r(h_t(p_i))$, both Bloom filters have the value 1. Although two Bloom filters both having 1 for all these r locations does not necessarily mean that $F(d_1)$ and $F(d_2)$ share a common The purpose of choosing r unique, if two Bloom filters have more 1s at the same locations than other pairs,



then the probability that they share common prefixes is higher.

Figure 5 shows the above hashing process for Bloom filters.

Finally, our privacy preserving index approach can be summarized as follows. Suppose that sensor S_i collects n_i data items in time slot t, in the end of this time slot, S_i executes the following six steps to encrypt the data items and build privacy preserving index. First, for each collected data item d_j ($1 \le j \le n_i$), S_i computes d_j for it according to the transformation table. Second, S_i runs the partition algorithm to partition each distinct data item d_j into a unique interval and embeds the interval information as $d_j|(b_j, e_j)$. Third, S_i encrypts d_j with $d_j|(b_j, e_j)$ into $E_i(d_j)$ using key k_i. Fourth, S_i computes d^m_j for d_j by choosing the lower bound of the interval containing d_j. Fifth, S_i creates a Bloom filter B_j for

storing d^m_j. Finally, S_i sends $(B_1, B_2, \cdots, B_{n_i})$ along with $(E_i(d_1), E_i(d_2), \cdots, E_i(d_{n_i}))$ to the nearby storage node.

A. Top-k to Top-Range Query

Our approach to executing top-k queries on the sensor data is to convert the top-k query into a *top-range* query. The intuition behind this is that, directly performing top-k queries on a set of sensor data items requires comparisons among them. But, given the prefix membership scheme from the previous section, we can check if a particular range of query prefixes are matched by any of the stored prefixes in the Bloom filter. This forms the basis of our approach wherein we transform the top-k query into a suitably crafted range query, which will obtain the same results as the top-k query. We next describe the details of this transformation.

JJJ. Now, for the approximated uniform distribution of the sensor values on $[d_0, d_{n+1}]$, the sink node needs to accurately estimate a range $[d_0, d]$,

which is equivalent to the top-k query over $[d_0, d_{n+1}]$. We call such ranges top-ranges as they always start from the lower bound d_0 . The value of d is important as it denotes the reference value for the top-k items to be retrieved. We note that, the probability that a data item falls
f (d) is the probability that the sink fails to get all k smallest data items in one top-range query with the range $[d_0, d]$. If we require $f(d) \leq c$, where c is suitably chosen constant, the lower bound of d can be obtained by solving equation (VI.2) for $f(d) = c$. However, since solving this equation analyt-ically is difficult we describe a binary search approach to compute the value of d .

Algorithm 1 Top-Range Computation

Input: $[d_0, d_{n+1}]$, n, k, c

Output: $[d_0, d]$

```
1  low := d0; high := dn+1 ;
2  while low ≤ high do
3      m = (low + high)/2;
4      p = (m − d0)/(dn+1 − d0);
5      f(d) = Σ_{j=0}^{k−1} C(n, j) p^j (1 − p)^{n−j};
6      if f(d) ≤ c then
7          high := m − 1;
8      else
9          low := m + 1;
10 d := low;
11 return [d0, d];
```

Based on the observation that $f(d)$ decreases with the increasing values of d , we use a binary search algorithm, shown in Algorithm 1, to compute d . The binary search narrows down the searching range for the value of d to one point. The starting value of d will be chosen from range
$[d_0, d_{n+1}]$. In the initial stage, we let $d = (d_0 + d_{n+1})/2$ and calculate $f(d)$ according to equation VI.2. If $f(d) \leq c$, we let the new range to be $[d_0, (d_0 + d_{n+1})/2 - 1]$ and d to be the middle value of this new range; Otherwise, we let the new range to be $[(d_0 + d_{n+1})/2 + 1, d_{n+1}]$ and d to be the middle value of this new range. This process continues until the value
of the low bound and the upper bound of a searching range become equal, and then the equal bound value is returned as the value for d .

When the sink wants to query k smallest data items, it executes the binary search algorithm to compute d and then sends the range $[d_0, d]$ instead of the numeric value of k to the storage node. In the end, the storage node returns a query result containing k data items to the sink. There are three possible cases. (1) $k = k$. In this case, the result is exactly the top-k query result that the sink needs. (2) $k > k$. In this case, the result contains some false positives, and the sink gets the
KKK.
this case, the sink needs to query for more data. The problem is reduced as to find out top-($k-k$) data items from $n-k$ data items, which uniformly spread in $[d + 1, d_{n+1}]$. Therefore, the
sink again executes binary search on input ($[d , d_{n+1}]$, $n - k$, $k - k$, c) to compute a new range $[d + 1, d]$, and performs
another range query.

LLL.

B. Trapdoor Computation

In this phase, we describe our trapdoor generation approach at the sink node to optimize the query results and reduce the false positives. To reduce false positives in query results, for a top-range query with range $[d_0, d]$, the sink executes the following three steps to compute the trapdoor denoted as $Q_{[d_0,d]}$. Firstly, the sink computes $S([d_0 , d]) = \{p_1, \cdots, p_z\}$ for the top-range $[d_0, d]$. Secondly, for each prefix p_i, $p_i \in S([d_0 , d])$,
if p_i has child-0 and child-1, the sink computes p_i^0 and p_i^1; otherwise the sink first computes p^*_i and p^-_i, which is obtained by replacing the first bit value of p_i with $*$and is impossible to be in any Bloom filters. The results are organized as a 3-tuple of the form (p_i , p_i^0, p_i^1) or (p_i, p_i^*, p^-_i). Finally, for
each 3-tuple (p_i , p_i^0, p_i^1) or (p_i , p_i^*, p^-_i), the sink computes $h_t(p_i)$, $h_t(p_i^0)$, $h_t(p_i^1)$ or $h_t(p_i)$, $h_t(p_i^*)$, $h_t(p^-_i)$. And then,
the sink organizes these hashes as a matrix $Q_{[d_0,d]}$. The trapdoor of p_i corresponds to the ith row of $Q_{[d_0,d]}$. After the trapdoor computation, the sink sends $Q_{[d_0,d]}$ to the storage nodes.

C. Query Execution

To describe the query execution at a storage node, we con-sider the scenario where a sensor S_i has submitted the Bloom
filter indexes, (B_1 , B_2, $\cdots$, B_{n_i}) and the encrypted data items, ($E_i(d_1)$, $E_i(d_2)$, $\cdots$, $E_i(d_{n_i})$), to the storage node.
First, the storage node generates r numbers $R_1, R_2, \cdots, R_r$ from $E_i(d_j)$ using the method from Section V-B. Second, the storage node checks whether there exists two hashes H_{xy} in a same row x($1 \leq x \leq z$, $1 \leq y \leq 3$) in $Q_{[a,b]}$ such that for every q($1 \leq q \leq r$), $B_i[h_q(H_{xy})]$:= 1. If this

condition holds, the storage node puts $E(d_j)$ into the query result. After performing the querying over all the indexes of all the sensors, the storage node transmits the query results to the sink.

D. Integrity Verification for Query Results

Finally, upon receiving a query result of the query trapdoor $Q_{[d_0,d]}$ from a storage node, the sink verifies the integrity for the query result as the follows. The sink divides the received data items into groups according to their source sensors. Next, the sink recovers the interval information for these data items and calculates the union of intervals for each group, denoted as R_i^U for the group of data items originates from the i-th sensor. Third, the sink tests all R_i^U for two things: (1) whether $R_i^U \cap [d_0, d]$ is a continuous range; (2) whether $[d_0, d] \subseteq R_i^U$. If either test fails for any R_i^U, based on the assumption that sensors are trusted, the sink can conclude that the storage node does not faithfully return all data items according to the query. The rationale behind this verification is: a complete query result must include every data item that its interval overlaps with $[d_0, d]$. We note that, the query result may include data items that their intervals do not overlap with $[d_0, d]$, due to the false positives of Bloom filters

E. False Positive Rate Analysis

As each index is represented by a Bloom filter, the query results inevitably have false positives. In our scheme, we choose the number of hash functions r to control the false positive rate. When $r = \frac{m}{n} \times \ln 2$, we have the false positive rate of a single bloom filter as

$$f = (1 - (1 - m^{-1})^{rn})^r \approx (1 - e^{-rn/m})^r \approx 0.6185^{m/n}.$$

F. VII. Security Analysis.

In this section, we prove that our proposed scheme is secure under the IND-CKA secure model. We use $SHA-1$ as the hash function in our scheme. $SHA-1$ is a keyed *pseudo-random* hash function whose output cannot be distinguished from a truly random function with non-negligible probability by a probabilistic polynomial time adversary [11]. According to [14], a searchable symmetric encryption scheme is secure if a probabilistic polynomial time adversary cannot distinguish between the output of a real index, which uses pseudo-random functions, and a simulated index, which uses truly random functions, with non-negligible probability. We first introduce some definitions from [14], and then construct a simulated index for our scheme and thereby, prove its security under the IND-CKA model.

History (H_l):

Let $D = \{d_1, d_2, \cdots, d_n\}$ be a data set, $Q_l = \{q_1, q_2, \cdots, q_l\}$ be l top-range queries. A l-query history over D is a tuple $H_l = \{D, Q_l\}$.

View (A_V):

A_V is the view of the adversary of the history H_l. It includes the secure Index I for D, the encrypted data set $EN_K(D) = \{EN_K(D_1), EN_K(D_2), \cdots, EN_K(D_n)\}$, and the trapdoors $T = \{t_{q_1}, t_{q_2}, \cdots, t_{q_l}\}$ for Q_l. The trapdoor for each top-range $q_i = [d_0, d_i]$ is a matrix t_{q_i}. Formally, $A_V =$

$$\{I, T, EN_K(D)\}.$$

VIII. Performance Evaluation

To evaluate the performance of our scheme, we considered two main parameters: one is the number of data items to be queried, denoted as n and another is the number of desired data items in the top-k query, *i.e.*, k. For various n and k, we comprehensively evaluated the top-k transformation method, the privacy preserving scheme, and the combination of these two. As we transform top-k queries to top-range queries, we compare our top-range query processing with the state-of-the-art range query processing schemes QuerySec [21] and SafeQ [19]. We use the average bandwidth cost, both for data submission from sensors to storage nodes and for query processing between storage nodes and the sink, as the performance metric.

A. Experimental Setup

We implemented our proposed scheme, QuerySec, and SafeQ using TOSSIM [36]. The implementations of QuerySec and SafeQ are according to [21] and [19], respectively. Our experimental configuration is as follows.

Data Set: We chose the same data set as used in [3], [19], and [21]. The data set was collected by Intel Lab [37],

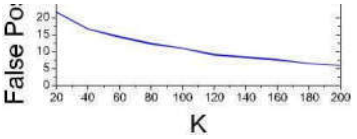


Fig. 6. False positive rates for range estimation strategy.

between 01/03/2004 to 03/10/2004, and consists of temper-ature, humidity and voltage data collected from 44 sensor nodes. We conducted our experiments on a single dimensional data (temperature). Similar to the methodology in QuerySec and SafeQ, we divided the 44 sensors into four groups and deployed a storage node in each group. We varied time slot period from 10 to 80 minutes to create different distributions on the number of data items that can be queried.

Top-k Query Transformation:

We first mapped the data items to a range [0, 10000] using the method introduced in Section IV-A. Then, we used our query transformation approach, from Section VI-A, to convert top-k queries to top-range queries. For each trial, considering that each group consists 11 sensors, we chose the values for k from 20 to 200. Our approach is simple but conservative, which results in high false positive rate in the query results. To reduce false positives, we adopted a *press on towards* method to estimate the range in experiments. We first used $k * 70\%$ to compute a range $[d_0, d_1]$ from $[d_0, d_{n+1}]$ and get k_1 data items. If $k_1 < k$, then we used $(k - k_1) * 90\%$ to compute a range $[d_1 + 1, d_2]$ from $[d_1, d_{n+1}]$ and got k_2 data items. If $k_1 + k_2 < k$, we used $(k - k_1 - k_2)$ to compute a range $[d_2 + 1, d_3]$ from $[d_3, d_{n+1}]$.

Finally, we evaluated the results based on the average query times and average false positive rates.

Privacy Preserving Index Generation:

We used *AES* cipher to encrypt data on every sensor, which is affordable with modern sensor hardware. We used SHA-1 as the pseudo-random one-way hash function function for the Bloom filter. We set the Bloom filter parameter $m/n = 8$, where m is the Bloom filter size and n is the number of elements, and the number of hash functions as 6. We used 14 bits to represent the data values in the [0, 10000]. Therefore, the number of possible prefixes in a prefix family of a data value is 15. The size of a Bloom filter is $15 * m/n$, *i.e.*, 15 bytes.

B. Summary for Experimental Results

We tested the efficiency of our optimization technique on the reduction of the false positive rates in top-k query to top-range query transformation. False positives in the query trans-formation cause data items below top-k to be transferred from storage nodes to the sink, and therefore, waste bandwidth. Note that, the false positive rate here is not the false positive rate introduced by Bloom filters. The experimental results show that our proposed scheme can control the false positive rates to a practically acceptable range. Figure 6 shows comparison of the false positive rate that we use our query transform approach to estimate top-range for a top-k query with *Press on Towards* strategy and the false positive rate that we use our

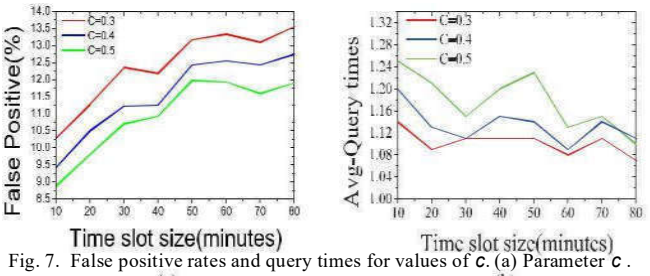


Fig. 7. False positive rates and query times for values of c. (a) Parameter c. (b) Query times.

Fig. 9. C

Fig. 8. Comparing bandwidth cost for sensor-storage nodes.

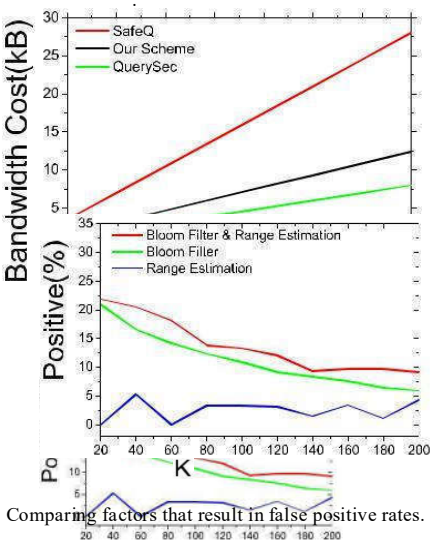
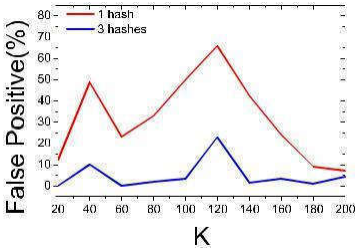


Fig. 11. Comparing factors that result in false positive rates.

9. Comparison for total bandwidth cost in three schemes. (a) $k = 100$. (b) $t = 40$ minutes.

Fig. 10.False positive rates from Bloom filter configurations.

C.



Comparison With Prior Art

We also compared our proposed scheme with two previ-ously proposed secure query processing schemes, QuerySec and SafeQ. The experimental results show that *our proposed scheme has practical accepted bandwidth cost.* Figure 8 shows the average bandwidth cost between a sensor and its nearby storage node in a time slot. The size of time slot varies from 10 minutes to 80 minutes. The experiment results indicate that our scheme consumes 0.55 times more bandwidth than QuerySec and spends 1.22 times less bandwidth of SafeQ. Figure 9 shows the average bandwidth cost of a top-k query between the sink and 4 storage nodes, using three different schemes. Figure 9(a) shows the experiment result that we fix $k = 100$ and time slot size ranging from 10 minutes to 80 minutes, and figure 9(b) shows the experiment result that we fix time slot $t = 40$ minutes and k ranging from 20 to 200. In both experiments, our proposed scheme consumes less bandwidth than the other two schemes in all tested cases.

IX. Conclusions

In this paper, we propose the first secure top-k query processing scheme that is secure under the IND-CKA secu-rity model. The data privacy is guaranteed by encryption as well as a careful generation of data indexes. We make two key contributions in this paper. The first contribution is to transform a top-k query to a top-range query and adopt membership testing to test whether a data item should be

included in the query result or not. This transformation allows the storage node to find k smallest or biggest data values without using numerical comparison operations, which is a key technique for the scheme to be secure under the IND-CKA security model. The second contribution is the data partition, index selection, and interval information embedding technique. This technique guarantees that at least one data item of each sensor collected data will be included in a query result and allows the sink to verify the integrity of query result without extra verification objects. Experiments show that the proposed scheme is bandwidth efficient and highly practical. The techniques proposed in this paper can be potentially useful for many other applications as well [38]–[53].

References

[1] P. Desnoyers, D. Ganesan, H. Li, M. Li, and P. Shenoy, "Presto:
A predictive storage architecture for sensor networks," in *Proc. 10th HotOS*, 2005, pp. 12–15.

[2] S. Ratnasamy *et al.*, "Data-centric storage in sensornets with GHT, a geographic hash table," *Mobile Netw. Appl.*, vol. 8, no. 4, pp. 427–442, Aug. 2003.

[3] B. Sheng and Q. Li, "Verifiable privacy-preserving range query in two-tiered sensor networks," in *Proc. 27th INFOCOM*, Apr. 2008, pp. 46–50.

[4] B. Sheng, Q. Li, and W. Mao, "Data storage placement in sensor networks," in *Proc. 7th ACM MobiHoc*, May 2006, pp. 344–355.

[5] D. Zeinalipour-Yazti, S. Lin, V. Kalogeraki, D. Gunopulos, and
W. A. Najjar, "Microhash: An efficient index structure for flash-based sensor devices," in *Proc. 4th USENIX FAST*, Dec. 2005, pp. 31–44.

[6] *Stargate Gateway (SPB400)*, accessed on 2011. [Online]. Available: http://www.xbow.com.

[7] *Rise Project*. accessed on 2011. [Online]. Available: http://www.cs.ucr.edu/~rise.

[8] I. F. Ilyas, G. Beskales, and M. A. Soliman, "A survey of top-k queries processing techniques in relational database system," *ACM Comput. Surv.*, vol. 40, no. 4, pp. 11:1–11:58, Oct. 2008.

[9] A. S. Silberstein, R. Braynard, C. Ellis, K. Munagala, and J. Yang, "A sampling-based approach to optimizing top-k queries in sensor networks," in *Proc. 22nd ICDE*, Apr. 2006, p. 68.

[10] R. Zhang, J. Shi, Y. Zhang, and X. Huang, "Secure top-k query processing in unattended tiered sensor networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4681–4693, Nov. 2014.

[11] E.-J. Goh, "Secure indexes," Cryptol. ePrint Archive, Rep. 2003/216. [Online]. Available: http://eprint.iacr.org/2003/216/

[12] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 222–233, Jan. 2014.

[13] B. Bezawada, A. X. Liu, B. Jayaraman, A. L. Wang, and R. Li, "Privacy preserving string matching for cloud computing," in *Proc. 35th ICDCS*, Jun./Jul. 2015, pp. 609–618.

[14] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: Improved definitions and efficient constructions,"
J. Comput. Secur., vol. 19, no. 5, pp. 895–934, Jan. 2011.

[15] P. Golle, J. Staddon, and B. Waters, "Secure conjunctive keyword search over encrypted data," in *Proc. 2nd ACNS*, 2004, pp. 31–45.

[16] O. Goldreich, *Foundations of Cryptography: Basic Tools*. Cambridge,
U.K.: Cambridge Univ. Press, 2001.

[17] X. Liao and J. Li, "Privacy-preserving and secure top-k query in two-tier wireless sensor network," in *Proc. 55th GLOBECOM*, Dec. 2012, pp. 335–341.

[18] C.-M. Yu, G.-K. Ni, I.-Y. Chen, E. Gelenbe, and S.-Y. Kuo, "Top-k query result completeness verification in tiered sensor networks," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 1, pp. 109–124, Jan. 2014.

[19] F. Chen and A. X. Liu, "SafeQ: Secure and efficient query processing in sensor networks," in *Proc. 29th INFOCOM*, Mar. 2010, pp. 2642–2650.

[20] J. Shi, R. Zhang, and Y. Zhang, "Secure range queries in tiered sensor networks," in *Proc. 28th INFOCOM*, Apr. 2009, pp. 945–953.

[21] Y. Yi, R. Li, F. Chen, A. X. Liu, and Y. Lin, "A digital watermark-ing approach to secure and precise range query processing in sensor networks," in *Proc. 32nd INFOCOM*, Apr. 2013, pp. 1998–2006.

[22] R. Zhang, J. Shi, and Y. Zhang, "Secure multidimensional range queries in sensor networks," in *Proc. 10th ACM MobiHoc*, May 2009, pp. 197–206.

[23] B. Hore, S. Mehrotra, and G. Tsudik, "A privacy-preserving index for range queries," in *Proc. 30th VLDB*, Aug. 2004, pp. 720–731.

[24] M. Naveed, S. Kamara, and C. V. Wright, "Inference attacks on property-preserving encrypted databases," in *Proc. 22nd ACM CCS*, Oct. 2015,
pp. 644–655.

[25] H. Hacigumus, B. Iyer, C. Li, and S. Mehrotra, "Executing SQL over encrypted data in the database-service-provider model," in *Proc. 21st ACM SIGMOD*, Jun. 2002, pp. 216–227.

[26] B. Hore, S. Mehrotra, M. Canim, and M. Kantarcioglu, "Secure multi-dimensional range queries over outsourced data," in *Proc. 21st VLDB*, Jun. 2012, pp. 333–358.

[27] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in *Proc. 23rd SIGMOD*, Jun. 2004,
pp. 563–574.

[28] A. Boldyreva, N. Chenette, Y. Lee, and A. O'Neill, "Order-preserving symmetric encryption," in *Proc. 23rd EUROCRYPT*, Apr. 2009,

[29] 224–241. A. Boldyreva, N. Chenette, and A. O'Neill, "Order-preserving encryp-tion revisited: Improved security analysis and alternative solutions," in *Proc. 31st CRYPTO*, Aug. 2011, pp. 578–595.

[30] R. Li, A. X. Liu, L. Wang, and B. Bezawada, "Fast range query processing with strong privacy protection for cloud computing," in *Proc. 40th VLDB*, Oct. 2014, pp. 1953–1964.

[31] D. Boneh and B. Waters, "Conjunctive, subset, and range queries on encrypted data," in *Proc. 4th Theory Cryptograph. Conf.*, Feb. 2007,
pp. 535–554.

[32] E. Shi, J. Bethencourt, T.-H. Chan, D. Song, and A. Perrig, "Multi-dimensional range query over encrypted data," in *Proc. 28th IEEE Symp. Secur. Privacy*, May 2007, pp. 350–364.

[33] A. X. Liu and F. Chen, "Collaborative enforcement of firewall policies in virtual private networks," in *Proc. 27th ACM PODC*, Aug. 2008,
pp. 95–104.

[34] B. H. Bloom, "Space/time tradeoffs in in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, Jul. 1970.

[35] Y.-C. Chang and M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in *Proc. 3rd ACNS*, Jun. 2005,
pp. 442–455.

[36] *Tossim*, accessed on 2011. [Online]. Available: http://www.cs.berkeley.edu/pal/research/tossim.html

[37] *Intel Lab Data*, accessed on 2011. [Online]. Available: http://berkeley.intel-research.net/labdata

[38] Z. Fu, F. Huang, X. Sun, A. Vasilakos, and C.-N. Yang, "Enabling semantic search based on conceptual graphs over encrypted out-sourced data," *IEEE Trans. Serv. Comput.*, to be published, doi: 10.1109/TSC.2016.2622697.

[39] Z. Fu, K. Ren, J. Shu, X. Sun, and F. Huang, "Enabling personalized search over encrypted outsourced data with efficiency improvement," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 9, pp. 2546–2559, Sep. 2016.

[40] Z. Fu, X. Sun, S. Ji, and G. Xie, "Towards efficient content-aware search over encrypted outsourced data in cloud," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.

[41] Z. Fu, X. Sun, Q. Liu, L. Zhou, and J. Shu, "Achieving efficient cloud search services: Multi-keyword ranked search over encrypted cloud data supporting parallel computing," *IEICE Trans. Commun.*, vol. E98B, no. 1, pp. 190–200, Jan. 2015.

[42] Z. Fu, X. Wu, C. Guan, X. Sun, and K. Ren, "Toward efficient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 12,
pp. 2706–2716, Dec. 2016.

[43] B. Gu and V. S. Sheng, "A robust regularization path algorithm for v-support vector classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 5, pp. 1241–1248, May 2017.

[44] Z. Xia *et al.*, "A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 11, pp. 2594–2608, Nov. 2016

[45] Z. Zhou, Y. Wang, Q. M. J. Wu, C.-N. Yang, and X. Sun, "Effective and efficient global context verification for image copy detection," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 48–63, Jan. 2017.

[46] Z. Xia, X. Wang, X. Sun, and Q. Wang, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data," *IEEE Trans. Parallel Distrib. Syst.*, vol 27, no. 2, pp. 340–352, Feb. 2016.

[47] Z. Qu, J. Keeney, S. Robitzsch, F. Zaman, and X. Wang, "Multilevel pattern mining architecture for automatic network monitoring in hetero-geneous wireless communication networks," *China Commun.*, vol. 13, no. 7, pp. 108–116, Jul. 2016.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Privacy Protection based Access Control Scheme in Cloud-based Services

SUPRIYA,
PG STUDENT,
Department of CSE
Mala Reddy College of Engineering
Hyderabad, India
Supriyachinna75@gmail.com

Dr.T. SUNIL
PROFESSOR
Department of CSE
Jawaharlal Nehru Technological University
Hyderabad, India.
sunil.tekale2010@gmail.com

Abstract— With the quick advancement of PC innovation, cloud-based administrations have turned into a hotly debated issue. They furnish clients with comfort, as well as bring numerous security issues, for example, information sharing and protection issue. In this paper, we show an entrance control framework with benefit detachment in view of security insurance (PS-ACS). In the PS-ACS plot, we isolate clients into a private area (PRD) and open space (PUD) legitimately. In PRD, to accomplish read get to authorization and compose get to consent, we embrace the Key-Aggregate Encryption (KAE) and the Improved Attribute-based Signature (IBS) separately. In PUD, we build another multi-specialist ciphertext approach quality based encryption (CP-ABE) conspire with productive decoding to stay away from the issues of single purpose of disappointment and entangled key conveyance, and plan a proficient property repudiation strategy for it. The investigation and reproduction result demonstrates that our plan is practical and better than ensure clients' security in cloud-based administrations.

Introduction (*Heading 1*)

With the rapid development of cloud computing, big data and public cloud services have been widely used. The user can store his data in the cloud service. Although cloud computing brings great convenience to enterprises and users, the cloud computing security has always been a major hazard. For users, it is necessary to take full advantage of cloud storage service, and also to ensure data privacy. Therefore, we need to develop an effective access control solution. Since the traditional access control strategy [1] cannot effectively solve the security problems that exist in data sharing. Data security issues brought by data sharing have seriously hindered the development of cloud computing, various solutions to achieve encryption and decryption of data sharing have been proposed. In 2007, Bethencourt et al. [2] first proposed the ciphertext policy attribute-based encryption (CP-ABE). However, this scheme does not consider the revocation of access permissions. In 2011, Hur et al. [3] put forward a fine-grained revocation scheme but it can easily cause key escrow issue. Lewko et al.

LITERATURE SURVEY

1. Cipher text-Policy Attribute-Based Encryption

AUTHORS: Taeho Jung1, Xiang-Yang Li12, Zhiguo Wan34, Meng Wan5

In several distributed systems a user should only be able to access data if a user posses a certain set of credentials or attributes. Currently, the only method for enforcing such policies is to employ a trusted server to store the data and mediate access control. However, if any server storing the data is compromised, then the confidentiality of the data will be compromised. In this paper we present a system for realizing complex access control on encrypted data that we call Ciphertext-Policy Attribute-Based Encryption. By using our techniques encrypted data can be kept confidential even if the storage server is untrusted; moreover, our methods are secure against collusion attacks. Previous AttributeBased Encryption systems used attributes to describe the encrypted data and built policies into user's keys; while in our system attributes are used to describe a user's credentials, and a party encrypting data determines a policy for who can decrypt. Thus, our methods are conceptually closer to traditional access control methods such as Role-Based Access Control (RBAC). In addition, we provide an implementation of our system and give performance measurements.

2. Multi-authority attribute based encryption with honest-but-curious central authority

AUTHORS: Vladimir Bo˘zovi´c1 , Daniel Socek2? , Rainer Steinwandt1 , and Vikt´oria I. Vill´anyi

An attribute based encryption scheme capable of handling multiple authorities was recently proposed by Chase. The scheme is built upon a single-authority attribute based encryption scheme presented earlier by Sahai and Waters. Chase's construction uses a trusted central authority that is inherently capable of decrypting arbitrary ciphertexts created within the system. We present a multi-authority attribute based encryption scheme in which only the set of recipients defined by the encrypting party can decrypt a corresponding ciphertext. The central authority is viewed as "honest-but-curious": on the one hand it honestly follows the protocol, and on the other hand it is curious to decrypt arbitrary ciphertexts thus violating the intent of the encrypting party. The

proposed scheme, which like its predecessors relies on the Bilinear DiffieHellman assumption, has a complexity comparable to that of Chase's scheme. We prove that our scheme is secure in the selective ID model and can tolerate an honest-but-curious central authority. Building on the proposal for multi-authority based attribute based encryption from [4], we constructed a scheme where the central authority is no longer capable of decrypting arbitrary ciphertexts created within the system. In addition to showing security in the selective ID model, we showed that the proposed system can tolerate an honest-but-curious central authority. Since both Chase's scheme and the proposed scheme rely on the same hardness assumption, and have a comparable complexity, the new scheme seems a viable alternative to Chase's construction. However, since only the proposed method is capable of handling a curious yet honest central authority, the proposed scheme is recommended in applications where security against such a central authority is required.

3. Decentralizing Attribute-Based Encryption

AUTHORS: Allison Lewko ∗ University of Texas at Austin alewko@cs.utexas.edu

We propose a Multi-Authority Attribute-Based Encryption (ABE) system. In our system, any party can become an authority and there is no requirement for any global coordination other than the creation of an initial set of common reference parameters. A party can simply act as an ABE authority by creating a public key and issuing private keys to different users that reflect their attributes. A user can encrypt data in terms of any boolean formula over attributes issued from any chosen set of authorities. Finally, our system does not require any central authority. In constructing our system, our largest technical hurdle is to make it collusion resistant. Prior Attribute-Based Encryption systems achieved collusion resistance when the ABE system authority "tied" together different components (representing different attributes) of a user's private key by randomizing the key. However, in our system each component will come from a potentially different authority, where we assume no coordination between such authorities. We create new techniques to tie key components together and prevent collusion attacks between users with different global identifiers. We prove our system secure using the recent dual system encryption methodology where the security proof works by first converting the challenge ciphertext and private keys to a semi-functional form and then arguing security. We follow a recent variant of the dual system proof technique due to Lewko and Waters and build our system using bilinear groups of composite order. We prove security under similar static assumptions to the LW paper in the random oracle model.

4. Accountable Authority Ciphertext-Policy Attribute-Based Encryption with White-Box Traceability and Public Auditing in the Cloud

AUTHORS: Jianting Ning1 , Xiaolei Dong2 , Zhenfu Cao2 and Lifei Wei3

As a sophisticated mechanism for secure fine-grained access control, ciphertext-policy attribute-based encryption (CP-ABE) is a highly promising solution for commercial applications such as cloud computing. However, there still exists one major issue awaiting to be solved, that is, the prevention of key abuse. Most of the existing CP-ABE systems missed this critical functionality, hindering the wide utilization and commercial application of CP-ABE systems to date. In this paper, we address two practical problems about the key abuse of CP-ABE: (1) The key escrow problem of the semi-trusted authority; and, (2) The malicious key delegation problem of the users. For the semi-trusted authority, its misbehavior (i.e., illegal key (re-)distribution) should be caught and prosecuted. And for a user, his/her malicious behavior (i.e., illegal key sharing) need be traced. We affirmatively solve these two key abuse problems by proposing the first accountable authority CP-ABE with whitebox traceability that supports policies expressed in any monotone access structures. Moreover, we provide an auditor to judge publicly whether a suspected user is guilty or is framed by the authority. In this work, we addressed two practical problems about the key abuse of CPABE in the cloud, and have presented an accountable authority CP-ABE system supporting white-box traceability and public auditing. Specifically, the proposed system could trace the malicious users for illegal key sharing. And for the semitrusted authority, its illegal key (re-)distributing misbehavior could be caught and prosecuted. Furthermore, we have provided an auditor to judge whether a malicious user is innocent or framed by the authority. As far as we known, this is the first CP-ABE system that simultaneously supports white-box traceability, accountable authority and public auditing. We have also proved that the new system is fully secure in the standard model. Note that there exists a stronger notion for traceability called black-box traceability. In black-box scenario, the malicious user could hide the decryption algorithm by tweaking it, as well as the decryption key. And in this case, the proposed system with white-box traceability in this paper will fail since both the decryption key and decryption algorithm are not well-formed. In our future work, we will focus on constructing an accountable

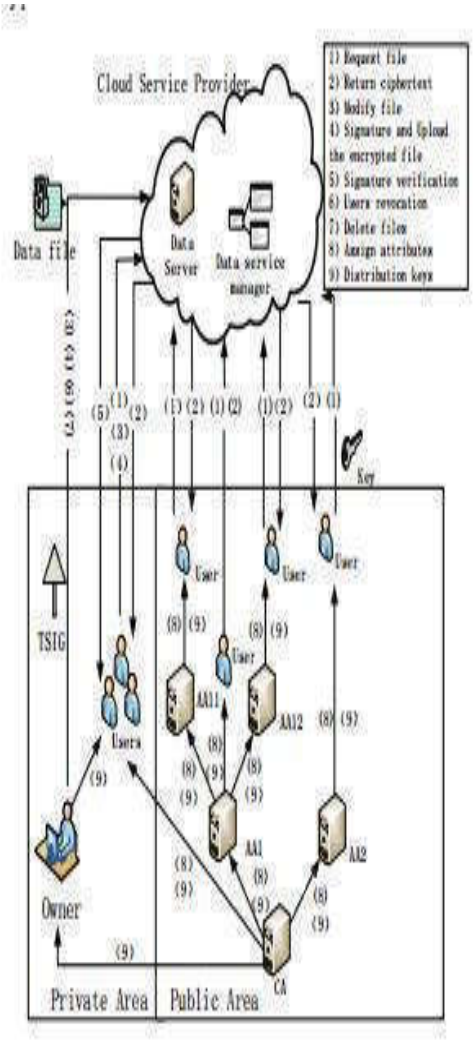
authority CP-ABE system which is black-box traceability and public auditing.

5.Privacy-preserving Attribute Based Searchable Encryption

AUTHORS: Payal Chaudhari1 and Manik Lal Das2

Attribute Based Encryption (ABE) is a promising public-key cryptographic primitive that can be used for cryptographically enforced access control in untrusted storage. Storing data on untrusted storage not only requires data security for data owners but also poses data protection from untrusted storage server. To address this important requirement, Anonymous Attribute Based Encryption (AABE) is a suitable primitive that provides users to access data from untrusted storage without revealing their identities. At the same time user data can be stored in untrusted storage in an encrypted form. While storing data in an encrypted form, keyword-based query search (and data retrieval) is a challenging research problem. In this paper we present an anonymous attribute based searchable encryption (A2SBE) scheme which facilitates user to retrieve only a subset of documents pertaining to his chosen keyword(s). User can upload documents in public cloud in an encrypted form, search documents based on keyword(s) and retrieve documents without revealing his identity. The scheme is proven secure under the selective ciphertextpolicy and chosen plaintext attack (IND-sCP-CPA) model and selective ciphertext-policy and chosen keyword attack (IND-sCP-CKA) model. The scheme requires small storage for user's decryption key and reduced computation for decryption in comparison to other schemes. The anonymous ABE provides an interesting security feature receiver anonymity in addition to data confidentiality and fine-grained access control of ABE. While storing encrypted documents in public cloud, an efficient search functionality facilitates user to retrieve a subset of documents for which the user has access rights on stored documents. We proposed an anonymous attribute based searchable encryption (A2SBE) scheme which facilitates user to retrieve only a subset of documents pertaining to his chosen keyword(s). User can upload documents in public cloud in an encrypted form, search documents based on keyword(s) and retrieve documents without revealing his identity. The scheme is proven secure under the standard adversarial model. The scheme is efficient, as it requires small storage for user's decryption key and reduced computation for decryption in comparison to other schemes.

ARCHITECTURE



REFERENCES

[1] A. Shamir, "Identity-based cryptosystems and signature schemes," in CRYPTO. Springer, 1985, pp. 47–53.

[2] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in EUROCRYPT. Springer, 2005, pp. 457–473.

[3] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in CCS. ACM, 2006, pp. 89–98.

[4] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attributebased encryption," in S&P. IEEE, 2007, pp. 321–334.

[5] M. Chase, "Multi-authority attribute based encryption," in TCC. Springer, 2007, pp. 515–534.

[6] M. Chase and S. S. Chow, "Improving privacy and security in multiauthority attribute-based encryption," in CCS. ACM, 2009, pp. 121– 130.

[7] H. Lin, Z. Cao, X. Liang, and J. Shao, "Secure threshold multi authority attribute based encryption without a central authority," Information Sciences, vol. 180, no. 13, pp. 2618–2632, 2010.

[8] V. Bozovi ˇ c, D. Socek, R. Steinwandt, and V. I. Vill ´ anyi, "Multi- ´ authority attribute-based encryption with honest-but-curious central authority," IJCM, vol. 89, no. 3, pp. 268–283, 2012.

[9] F. Li, Y. Rahulamathavan, M. Rajarajan, and R.-W. Phan, "Low complexity multi-authority attribute based encryption scheme for mobile cloud computing," in SOSE. IEEE, 2013, pp. 573–577.

[10] K. Yang, X. Jia, K. Ren, and B. Zhang, "Dac-macs: Effective data access control for multi-authority cloud storage systems," in INFOCOM. IEEE, 2013, pp. 2895–2903.

[11] A. Lewko and B. Waters, "Decentralizing attribute-based encryption," in EUROCRYPT. Springer, 2011, pp. 568–588.

[12] S. Muller, S. Katzenbeisser, and C. Eckert, "On multi-authority ¨ ciphertext-policy attribute-based encryption," Bulletin of the Korean Mathematical Society, vol. 46, no. 4, pp. 803–819, 2009.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

SENTIMENT EMBEDDINGS WITH APPLICATINS TO APPLICATIN SENTIMENT ANALYSIS

G.Priskilla,
PG STUDENT,
Department of CSE
Jawaharlal Nehru Technological University
Hyderabad, India.
lovelyprisk514@gmail.com

Dr.T. Sunil
PROFESSOR
Department of CSE
Mala Reddy College of Engineering
Hyderabad, India
sunil.tekale2010@gmail.com

Abstract— **We propose learning sentiment-specific word embeddings dubbed sentiment embeddings in this paper. Existing word embedding learning algorithms typically only use the contexts of words but ignore the sentiment of texts. It is problematic for sentiment analysis because the words with similar contexts but opposite sentiment polarity, such as good and bad, are mapped to neighboring word vectors. We address this issue by encoding sentiment information of texts (e.g., sentences and words) together with contexts of words in sentiment embeddings. By combining context and sentiment level evidences, the nearest neighbors in sentiment embedding space are semantically similar and it favors words with the same sentiment polarity. In order to learn sentiment embeddings effectively, we develop a number of neural networks with tailoring loss functions, and collect massive texts automatically with sentiment signals like emoticons as the training data. Sentiment embeddings can be naturally used as word features for a variety of sentiment analysis tasks without feature engineering. We apply sentiment embeddings to word-level sentiment analysis, sentence level sentiment classification, and building sentiment lexicons. Experimental results show that sentiment embeddings consistently outperform context-based embeddings on several benchmark datasets of these tasks. This work provides insights on the design of neural networks for learning task-specific word embeddings in other natural language processing tasks.**

Introduction (*Heading 1*)

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

How Data Mining Works?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks.

LITERATURE SURVEY

1) Learning Sentiment-specific word embedding for twitter sentiment classification

AUTHORS: D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B.

We present a method that learns word embedding for Twitter sentiment classification in this paper. Most existing algorithms for learning continuous word representations typically only model the syntactic context of words but ignore the sentiment of text. This is problematic for sentiment analysis as they usually map words with similar syntactic context but opposite sentiment polarity, such as good and bad, to neighboring word vectors. We address this issue by learning sentimentspecific word embedding (SSWE), which encodes sentiment information in the continuous representation of words. Specifically, we develop three neural networks to effectively incorporate the supervision from sentiment polarity of

text (e.g. sentences or tweets) in their loss functions. To obtain large scale training corpora, we learn the sentiment-specific word embedding from massive distant-supervised tweets collected by positive and negative emoticons. Experiments on applying SSWE to a benchmark Twitter sentiment classification dataset in SemEval 2013 show that (1) the SSWE feature performs comparably with hand-crafted features in the top-performed system; (2) the performance is further improved by concatenating SSWE with existing feature set.

2. Coooolll: A deep learning system for twitter sentiment classification

AUTHORS: D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou

In this paper, we develop a deep learning system for message-level Twitter sentiment classification. Among the 45 submitted systems including the SemEval 2013 participants, our system (Coooolll) is ranked 2nd on the Twitter2014 test set of SemEval 2014 Task 9. Coooolll is built in a supervised learning framework by concatenating the sentiment-specific word embedding (SSWE) features with the state-of-the-art hand-crafted features. We develop a neural network with hybrid loss function 1 to learn SSWE, which encodes the sentiment information of tweets in the continuous representation of words. To obtain large-scale training corpora, we train SSWE from 10M tweets collected by positive and negative emoticons, without any manual annotation. Our system can be easily re-implemented with the publicly available sentiment-specific word embedding.

3. A neural probabilistic language model

AUTHORS: Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin

A goal of statistical language modeling is to learn the joint probability function of sequences of words in a language. This is intrinsically difficult because of the **curse of dimensionality**: a word sequence on which the model will be tested is likely to be different from all the word sequences seen during training. Traditional but very successful approaches based on n-grams obtain generalization by concatenating very short overlapping sequences seen in the training set. We propose to fight the curse of dimensionality by **learning a distributed representation for words** which allows each training sentence to inform the model about an exponential number of semantically neighboring sentences. The model learns simultaneously (1) a distributed representation for each word along with (2) the probability function for word sequences, expressed in terms of these representations. Generalization is obtained because a sequence of words that has never been seen before gets high probability if it is made of words that are similar (in the sense of having a nearby representation) to words forming an already seen sentence. Training such large models (with millions of parameters) within a reasonable time is itself a significant challenge. We report on experiments using neural networks for the probability function, showing on two text corpora that the proposed approach significantly improves on state-of-the-art n-gram models, and that the proposed approach allows to take advantage of longer contexts.

4) Distributed representations of words and phrases and their compositionality

AUTHORS: T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling.

An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of "Canada" and "Air" cannot be easily combined to obtain "Air Canada". Motivated by this example, we present a simple method for finding phrases in text, and show that learning good vector representations for millions of phrases is possible

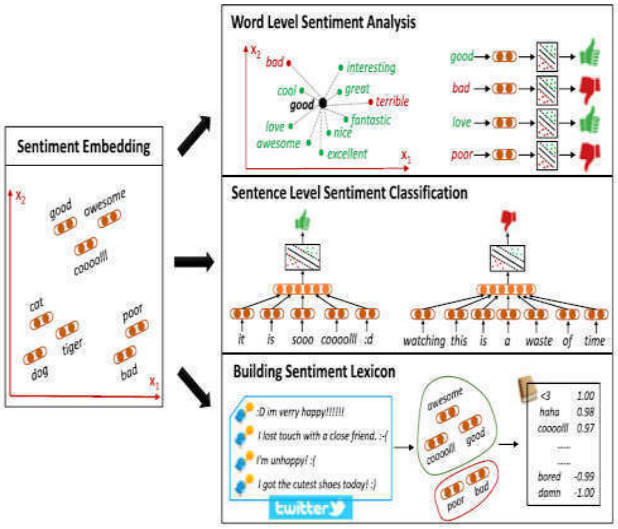
5) Word alignment modeling with context dependent deep neural network

AUTHORS: N. Yang, S. Liu, M. Li, M. Zhou, and N. Yu

In this paper, we explore a novel bilingual word alignment approach based on DNN (Deep Neural Network), which has been proven to be very effective in various machine learning tasks (Collobert et al., 2011). We describe in detail how we adapt and extend the CD-DNNHMM (Dahl et al., 2012) method introduced in speech recognition to the HMMbased word alignment model, in which bilingual word embedding is discriminatively learnt to

capture lexical translation information, and surrounding words are leveraged to model context information in bilingual sentences. While being capable to model the rich bilingual correspondence, our method generates a very compact model with much fewer parameters. Experiments on a large scale English- Chinese word alignment task show that the proposed method outperforms the HMM and IBM model 4 baselines by 2 points in F-score.

ARCHITECTURE



.References

[1] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning Sentiment-specific word embedding for twitter sentiment classification," in Proc. 52th Annu. Meeting Assoc. Comput. Linguistics., 2014, pp. 1555–1565.

[2] D. Tang, F. Wei, B. Qin, M. Zhou, and T. Liu, "Building large-scale twitter-specific sentiment lexicon: A representation learning approach," in Proc. 25th Int. Conf. Comput. Linguistics, 2014, pp. 172–182.

[3] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Coooolll: A deep learning system for twitter sentiment classification," in Proc. 8th Int. Workshop Semantic Eval., 2014, pp. 208–212.

[4] C. D. Manning and H. Sch€utze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[5] D. Jurafsky and H. James, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Englewood Cliffs, NJ, USA: Prentice-Hall, 2000.

[6] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," J. Mach. Learning Res., vol. 3, pp. 1137–1155, 2003.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proc. Conf. Neural Inf. Process. Syst., 2013, pp. 3111–3119.

[8] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proc. Conf. Empirical Methods Natural Lang. Process., 2014, pp. 1532–1543.

[9] Z. S. Harris, "Distributional structure," Word, vol. 10, pp. 146–162, 1954.

[10] N. Yang, S. Liu, M. Li, M. Zhou, and N. Yu, "Word alignment modeling with context dependent deep neural network," in Proc. 51st Annu. Meeting Assoc. Comput. Linguistics, 2013, pp. 166–175.

[11] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," in Proc. Annu. Meeting Assoc. Comput. Linguistics, 2013, pp. 455–465.

[12] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. Daum_e III, "A neural network for factoid question answering over paragraphs," in Proc. Conf. Empirical Methods Natural Lang. Process., 2014, pp. 633–644.

[13] J. Li, R. Li, and E. Hovy, "Recursive deep models for discourse parsing," in Proc. Conf. Empirical Methods Natural Lang. Process., 2014, pp. 2061–2069.

[14] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," J. Mach. Learning Res., vol. 12, pp. 2493–2537, 2011.

[15] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retrieval, vol. 2, nos. 1/2, pp. 1–135, 2008.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Multi Disease Prediction using Data Mining Techniques

K.Ashok,
M.Tech [CSE] Scholar,
Malla Reddy College of Engineering

M. Mounika,
M.Tech [CSE] Scholar,
Malla Reddy College of Engineering

S.Rasheeduddin,
Asst.Prof. CSE Dept.
Malla Reddy College of Engineering

ABSTRACT

Data mining techniques are used for a variety of applications. In healthcare industry, data mining plays an important role in predicting diseases. For detecting a disease number of tests should be required from the patient. But using data mining technique the number of tests can be reduced. This reduced test plays an important role in time and performance. This paper analyzes data mining techniques which can be used for predicting different types of diseases. This paper reviewed the research papers which mainly concentrate on predicting heart disease, Diabetes and Breast cancer etc.

Keywords: Data Mining, Classification, Naive Bayes, J48, Decision tree.

INTRODUCTION

Data mining is the process of selecting, discovering and modeling huge amounts of data. This process has become an increasingly insidious activity in all areas of medical science research. Data mining has resulted in the discovery of useful hidden patterns from huge databases. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, soft computing and data visualization; and includes classification and regression techniques. Some of the research works are done in this side, but all of them are focusing on a few methods of analysis, diagnosis or prediction of this disease by using different tools and techniques, this work is focused on the early prediction of various diseases by using WEKA tool.

HEART DISEASE

Heart disease is the leading cause of death in the U.S. At any point in your life, either you or one of your loved ones will be forced to make decisions about some aspect of heart disease. Knowing about the structure and functioning of the heart, in particular how angina and heart attacks work, will enable to make informed decisions about your health. Heart disease can strike suddenly and need to make decisions quickly.

Several factors contribute to this damage. They include:

1. Smoking, including secondhand smoke
2. High amounts of certain fats and cholesterol in the blood
3. High blood pressure
4. High amounts of sugar in the blood due to insulin resistance or diabetes
5. Blood vessel inflammation

DIABETICS

Diabetes mellitus, or simply diabetes, is a chronic disease that occurs when the pancreas is no longer able to make insulin, or when the body cannot make good use of the insulin it produces. Insulin is a hormone made by

the pancreas, which acts like a key to let glucose from the food we eat pass from the blood stream into the cells in the body to produce energy. All carbohydrate foods are broken down into glucose in the blood. Insulin helps glucose get into the cells.

Not being able to produce insulin or use it effectively leads to raised glucose levels in the blood (known as hyperglycemia). Over the long-term high glucose levels are associated with damage to the body and failure of various organs and tissues.

Fast facts on diabetes

1. Diabetes is a long-term condition that causes high blood sugar levels.
2. In 2013 it was estimated that over 382 million people throughout the world had diabetes (Williams's textbook of endocrinology).
3. Type 1 Diabetes - the body does not produce insulin. Approximately 10% of all diabetes cases are type 1.
4. Type 2 Diabetes - the body does not produce enough insulin for proper function. Approximately 90% of all cases of diabetes worldwide are of this type.
5. Gestational Diabetes - this type affects females during pregnancy.
6. The most common diabetes symptoms include frequent urination, intense thirst and hunger, weight gain, unusual weight loss, fatigue, cuts and bruises that do not heal, male sexual dysfunction, numbness and tingling in hands and feet.
7. If you have Type 1 and follow a healthy eating plan, do adequate exercise, and take insulin, you can lead a normal life.
8. Type 2 patients need to eat healthily, be physically active, and test their blood glucose. They may also need to take oral medication, and/or insulin to con- trol blood glucose levels.
9. As the risk of cardiovascular disease is much higher for a diabetic, it is crucial that blood pressure and cholesterol levels are monitored regularly.
10. As smoking might have a serious effect on cardio- vascular health, diabetics should stop smoking.
11. Hypoglycemia - low blood glucose - can have a bad effect on the patient. Hyperglycemia - when blood glucose is too high - can also have a bad effect on the patient.

BREAST CANCER

Breast cancer is a malignant tumor that starts in the cells of the breast. A malignant tumor is a group of cancer cells that can grow up into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. The disease occurs almost entirely in women, but men can get it, too.

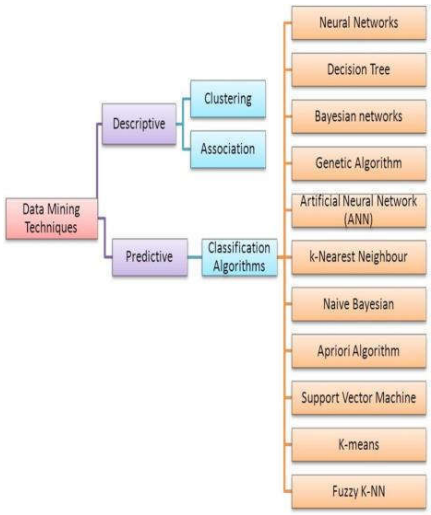
Breast cancer facts

1. Breast cancer is the most common cancer among American women.
2. One in every eight women in the United States develops breast cancer.
3. There are many types of breast cancer that differ in their capability of spreading (metastasize) to other body tissues.
4. The causes of breast cancer are not yet fully known, although a number of risk factors have been identified.
5. There are many different types of breast cancer.
6. Breast cancer symptoms and signs include
   a lump in the breast or armpit,
   bloody nipple discharge,
   inverted nipple,
   orange-peel texture or dimpling of the breast's skin,
   breast pain or sore nipple,
   swollen lymph nodes in the neck or armpit,
   Change in the size or shape of the breast or nipple.
   Breast cancer is diagnosed during a physical exam, by self-examination of the breasts, mammography, ultrasound testing, and biopsy.
   Treatment of breast cancer depends on the type of cancer and its stage (0-IV) and may involve surgery, radiation, or chemotherapy.

Data Mining Techniques

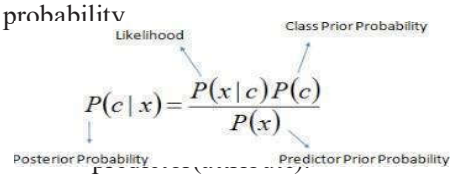
NAIVE BAYES

ISBN 978-81-939386-1-4



The Naive Bayesian classifier is based on Baye's theorem with independence assumptions between predictors. A Naive Bayesian model is simple to build, with no difficult iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Bayes theorem provides a way of manipulative the posterior probability



of class(target) given

2) P(c)is the prior probability of class.
3) P (x|c) is the likelihood which is the probability of predictor given class.
4) P(x)is the prior probability of predictor.

DECISION TREE (J48)

Decision tree models are commonly used in data mining to examine data and induce the tree and its rules that will be used to make predictions[10].The prediction could be to predict categorical values (classification trees) when instances are to be placed in categories or classes. Decision tree is a classifier in the structure of a tree structure where each node is either a leaf node, indicating the value of the target attribute or class of the examples, or a decision node, specifying some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. A decision tree can be used to classify an case by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance.

APPLICATION OF DATA MINING TECHNIQUES IN HEALTH CARE

The different classification algorithms mentioned below in figure 1 are used to predict or to analyze various diseases.

Figure 2. Different techniques in Healthcare domain

Summary of Techniques for Medical data mining.

In terms of prediction and decision making, Data mining techniques have substantial expansion in medical industry with respect to various diseases like diabetes, heart disease, liver diseases, cancer and others. Table 1 summarizes the medical data mining, its techniques used and for the related disease.

Table 1. Summary of medical data mining techniques

Disease	Technique Used
Conventional PathologyData	Extracting patterns & detecting trends using Neural Networks [3].
Coronary heart disease	Prediction models using Decision Tree Algorithms such as ID3, C4.5, C5, and CART [3].
Lymphoma Disease and Lung Cancer	Distinguish disease subtypes using Ensemble approach [4] [6].
Psychiatric Diseases	Predicate the probability of a psychiatric patient on the basis detected symptoms using BBN Bayesian networks [7].
Frequent Disease	Identify frequency of diseases in particular geographical using Apriori algorithm [8].
Liver diseases	Classification using Bayesian Ying Yang (BYY) [10].
Skin Disease	Categorization of skin disease using integrated decision model with neural network classification methods [14].
Diabetes	Classification of Medical Data using Genetic Algorithm [15].
Functional Magnetic Resonance Imaging (fMRI)	Integration of Clustering and Classification of biomedical databases [16].

Chest Disease	Constructed a model using Artificial Neural Network (ANN)
Diabetes, Cancer	Classification of Disease using k-Nearest Neighbor
Coronary Heart Disease	Improving classification accuracy using Naive Bayesian
Chronic Disease	Prediction of Diseases Using Apriori Algorithm
Diabetes	Disease classification using Support Vector Machine
Breast Cancer	Accurate Classification of medical data using K- means, Organizing Map (SOM) and Naïve Bayes
Cardio Vascular Diseases	Diagnose Cardio Vascular Disease using Classification algorithm.
Parkinson Disease	Familiarized an adaptive Fuzzy K-NN approach for diagnosing the disease

EXPERIMENTAL RESULTS

In this study, the accuracy of three data mining techniques is compared. The goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics such as specificity and sensitivity. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

In this study two different data mining classification techniques was used for the prediction of various diseases and their performance was compared in order to evaluate the best classifier. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications.

CONCLUSION

With the recent rapid rise in the quantity of biomedical data that is gathered by electronic means in critical care and the rampant availability of inexpensive and dependable computing equipment, many researchers has started, or are eager to start, exploring these data. In this paper we observe some data mining techniques that has been employed for medical data. As there is voluminous records in this industry and because of this, it has become requisite to use data mining techniques to help in decision support and prediction in the field of Healthcare to identify the kind of disease. The medical data mining produces business intelligence which is useful for diagnosing of the disease. This paper throws light into data mining techniques that is used for medical data for various diseases which are identified and diagnosed for human health.

Technique	*Diseases*	*Accuracy*
Naïve Bayes	Heart Disease	79%
	Diabetes	77.6%
	Breast Cancer	82.5%
J48	Heart Disease	77.033%
	Diabetes	100%
	Breast Cancer	75.52%

REFERENCES

1. http://www.mayoclinic.org/diseases-conditions/heart-disease/basics/definition/con-20034056.
2. http://www.idf.org/about-diabetes.
3. http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-what-is-breast-cancer.
4. Kumara, M., Vohra, R., Arora, A. (2014). Prediction of diabetes using Bayesian network. *International Journal of Computer Science and Information Technologies*, 5(4), 5174-5178.
5. Thirumal, P.C., & Nagarajan, N. (2015). Utilization of data mining techniques for diagnosis of diabetes mellitus - A case study. *ARPN Journal of Engineering and Applied Sciences*, January, 10(1), 8-13.
6. Gomathi, K. (2012). An empirical study on breast can- cer using data mining techniques. *International Journal of Research in Computer Application & Management*, July, 2(7), 97-102.

7. Witten, H. I., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, (2nd Ed), Morgan Kaufmann Publishers.
8. Witten, I. H. & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques, (2ndEd.) San Fransisco: Morgan Kaufmann.
9. WEKA: Data Mining Software in Java. Retrieved from
10. http://www.cs.waikato.ac.nz/ml/weka/.
11. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, June, 34(2), 113-127.
12. C. Hattice & K. Metin, "A DIAGNOSTIC SOFTWARE TOOL FOR SKIN DISEASES WITH BASIC AND WEIGHTED K-NN", Innovations in Intelligent Systems and Applications (INISTA), 2012.
13. Dhanya P Varghese & Tintu P B, "A SURVEY ON HEALTH DATA USING DATA MINING TECHNIQUES", International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 07, Oct-2015.
14. Doron Shalvi & Nicholas DeClaris, "AN UNSUPERVISED NEURAL NETWORK APPROACH TO MEDICAL DATA MINING TECHNIQUES", IEEE, 1998.
15. Gustavo Santos-Garcia & Gonzalo Varela & Nuria Novoa & Marcelo F. Jimenez, "PREDICTION OF POSTOPERATIVE MORBIDITY AFTER LUNG RESECTION USING AN ARTIFICIAL NEURAL NETWORK ENSEMBLE", Artificial Intelligence in Medicine 30:61–69, 2004.
16. Harleen Kaur & Siri Krishan Wasan, "EMPIRICAL STUDY ON APPLICATIONS OF DATA MINING TECHNIQUES IN HEALTHCARE", Journal of Computer Science 2 (2): 194-200, 2006.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Energy Efficient Resource Allocation in Cloud Computing

Dr. T Sunil
Professor
Dept. CSE
MRCE , Hyderabad, India.
Sunil.tekale2010@gmail.com
Ph : 9885303678

Mr. M Aharonu
Assistant Professor
Dept. CSE
MRCE , Hyderabad, India.
aharon.mattakoyya@gmail.com
ph: 9603031640

Mrs . Ch VijayaKumari
Associate Professor
Dept. CSE
MRCE , Hyderabad, India.
nagvijji@gmail.com
ph:9985855402

Abstract:

Cloud service provider in cloud environment will provide or provision resource based on demand from the user. The cloud service provider (CSP) will provide resources as and when required or demanded by the user for execution of the job on the cloud environment. The CSP will perform this in a static and dynamic manner. The CSP should also consider various other factors in order to provide the resources to the user, the prime among that will be the Service Level Agreement (SLA), which is normally signed by the user and cloud service provider during the inception phase of service. There are many algorithm which are used in order to allocate resources to the user in cloud environment. The algorithm which is proposed will be used to reduce the amount of energy utilized in performing various job execution in cloud environment. Here the energy utilized for execution of various jobs are taken into account by increasing the number of virtual machines that are used on a single physical host system. There is no thumb rule to calculate the number of virtual machines to be executed on a single host. The same can be derived by calculating the amount of space, speed required along with the time to execute the job on a virtual machine. Based up on this we can derive the number of

Virtual machine on a single host system. There can be 10 virtual machines on a single system or even 20 number of virtual machines on single physical system. But if the same is calculated by the equation then the result will be exactly matching with the threshold capacity of the physical system. If more number of physical systems are used to execute fewer virtual machines on each then the amount of energy consumed will be very high. So in order to reduce the energy consumption , the algorithm can be used will not only will help to calculate the number of virtual machines on single physical system , but also will help to reduce the energy as less number of physical systems will be in need.

Keywords: Virtual, machine, energy, physical, host, system, SLA.

Introduction:

Usage of various services such as software service, platform services and infrastructure service or hardware service in order to make an application run on internet from an external

source can be refereed as Cloud computing as in Fig: 2. The user here can makes use of these resources and can pay to the cloud service provider as per use. It is like a utility service such as an electricity bill and telephone bill as shown in Fig: 1 . The user here can enter into an agreement with the cloud service provider to make use of resource as and when required.

This agreement is referred as Service Level Agreement (SLA).



Fig:1



Fig:2

Virtualization is a process where on a physical system a virtual environment can be created so as to run a desired program by the user as in Fig:4. We can use this concept of virtualization for hardware virtualization, software and network virtualization. Coming to hardware virtualization there can be many subtypes in using this concept such as Full virtualization, Para virtualization and Emulation virtualization. Likewise subtypes also exists in software virtualization such as virtualization in Service called as service virtualization, Application virtualization and Operating system virtualization as in Fig: 3. In memory virtualization there can be two subtypes like Application level control and operating system level control. Further storage virtualization can also be at block level virtualization and file level virtualization.

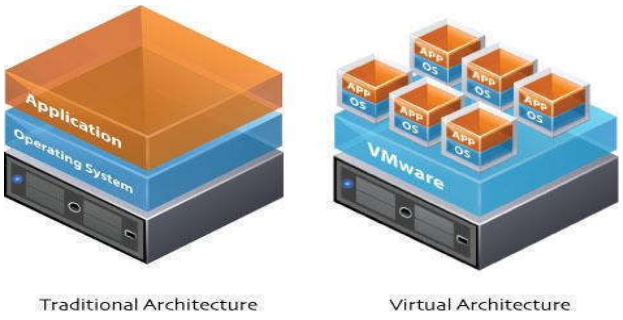


Fig:3

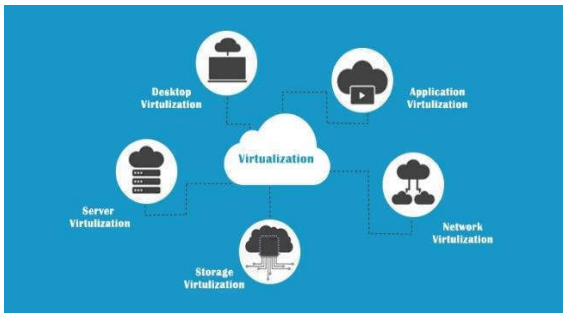


Fig:4

Access to the virtual machine and the host machine/ server is facilitated by a software known as Hypervisor. Hypervisor acts as a link between the hardware and the virtual environment and distributes the hardware resources such as CPU usage, memory allotment between the different virtual environments.

ISBN 978-81-939386-1-4

Related Work:

We normally use cloud services for storing the data, photos, and videos, documents using Amazon services or Microsoft Services or even google services for that matter. When cloud services are used infrequently the people performing research found that the power consumed is far less when it is compared with the power consumption in conventional computing and storage. Researchers found that the at low levels of usage the power consumption for storage of data and other things will dominate the total consumption of power as the power for transport is minimal or minimum. This is not true when the usage levels are at medium and high levels, more amount of energy is required to transmit data and because of which total consumption of power increases drastically over the overall energy consumed.

More precisely the power consumption for transmission of data can be very low for low usage of levels for public and private storage which can be 10% to 25% and the same will increase drastically for high usage levels and will be around 60% and 90%.

From this it is very clear that the cloud services when utilized for storage will use less amount of energy compared to cloud usage for processing of data and used for cloud computing. As long as the screen refresh rates are very low like say 0.1 frames/sec the amount of power consumption in transport is small in Cloud software services such as Google Documents, where as one 1 frame/sec means that 100% changes happens on the screen every second and so. Smaller the percentage change in the screen will reflect in low or smaller refresh rate. Biggest or important factor for the cloud software services where energy is to be determined working on a particular server. So in case like this the more beneficial one will be the public cloud users when compared with private cloud users because of its larger number of users working.

For cloud handling administrations (in which a server, for example, Amazon Elastic Compute Cloud forms extensive computational assignments just, and littler undertakings are prepared on the client's PC), the scientists again discovered that the cloud option can utilize bring down utilization just under specific conditions. The outcomes demonstrated that, for open cloud preparing administrations, information transport devoured a lot of vitality contrasted with private cloud handling administrations, especially at high utilization levels. The reason is that the vast number of switch bounces required on the general population Internet incredibly builds the vitality utilization in transport, and private cloud preparing requires fundamentally less switches. All things considered, the analysts found that, for both open and private mists, a cloud preparing administration is more vitality proficient than more seasoned age PCs. The overall results specify different concept for different type of users. As the research performed by the group states that the computer user at home can achieve significant amount of energy savings by making use of low end computers or laptops for performing the task which are of

ISBN 978-81-939386-1-4

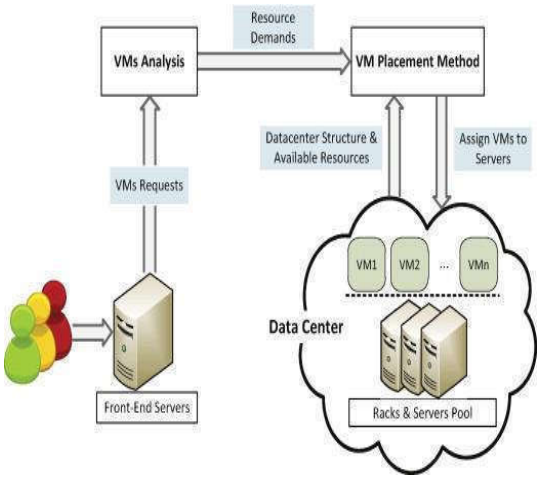
routine nature where as to perform the task occasionally which are of high computational nature the use of cloud service can be made use of instead of making use of high end PC or a mid-range PC.

For companies, it is less certain whether the vitality utilization spared in transport with a private cloud contrasted with an open cloud counterbalances the private cloud's higher vitality utilization. Private mists that serve a generally modest number of clients may not profit by a similar vitality sparing methods because of their littler scale.

As distributed computing turns out to be progressively across the board it is trusted that or expectations are to such an extent that the distributed computing will turn out to be increasingly more vitality proficient in information transport, particularly with the server farms where gadgets like ,switches ,centre points are utilized. As we know that the device which are used to transfer data from one place to another will have less or limited capacity, it is then the only solution available for the user to have data mobility is by making use of cloud services. Which most of the industry people believe and started working using cloud computing concepts.

The client gadget will contain minimal in excess of a program when it is begun up. Any application or information that it requires will be brought down from the cloud. At the point when that application is done, its information will be returned to the cloud and the application will be expelled from the client gadget until the point that it is again required. Along these lines, the client gadget is kept basic, vitality effective and shoddy." Various parts of asset the executives influence server control utilization? Understanding the impact of various parameters on power utilization is essential so as to oversee servers in a vitality productive way. These parameters may incorporate differing the quantity of virtual machines (VMs), fluctuating the figuring assets allotted to a VM, (for example, CPU centres and memory), and turning centres on/off, shifting the CPU recurrence and mapping virtual CPUs to physical CPU centres.

Proposed Algorithm:



Here we have consider the fact that the number of physical systems used will be directly proportional to the number of physical servers reaching the threshold capacity. i.e. to say if the number of physical systems used are ten, it indicates that the nine physical systems have reached the threshold capacity so the need for usage of tenth system has arised. If we consider the V1, V2, and V3 .... As the virtual machines and TrC as the threshold capacity of the physical system then

$\sum V1, V2, V3 \ldots Vn <= TrC$ and if the

$\sum V1, V2, V3 \ldots Vn >= TrC$ then it is

an indication that the next physical system has to be used for execution of various virtual machines.
If *P1, P2, P3* … are the physical systems and the power utilised by *P1, P2*… is *Po1, Po2* and so on then the number of systems reduced will be equal to the amount of power consumption reduction in the data centre.

Algorithm:

Begin

for each job in incoming queue
allocate virtual machine
if the sum of vms >=TrC
allocate VM on new physical system
else
allocate VM on existing physical system
for-each

Experimental Results:

$TrC\ S1=M(v1+v2+v3+v4)$

$TrC\ S2=M(v1+v2+v3+v4)$

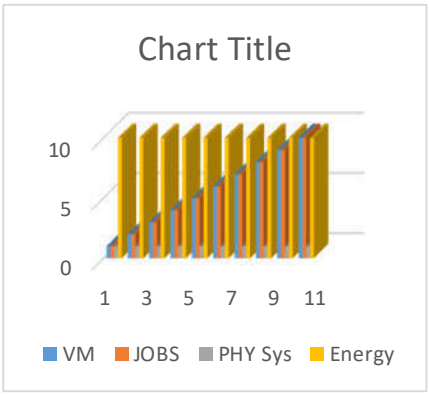
$TrC\ S3=M(v1+v2+v3+v4)$

$TrC\ S4=M(v1+v2+v3+v4)$

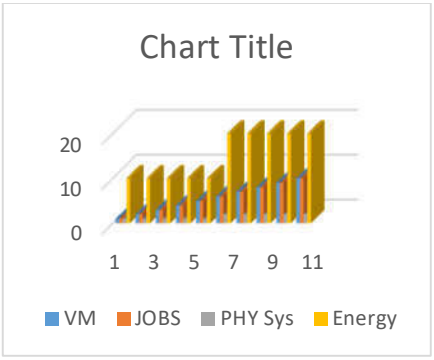
Threshold capacity of the System one is calculated as maximum capacity of virtual machine1 , virtual machine2 virtual machine3, virtual machine4.

So if S4 is started without the threshold of S3 is reached then it will consume more energy than what is required.Which means the enery consumed per se is 1 unit per system then we have to consume 4 units intead of three units to complete the task given.

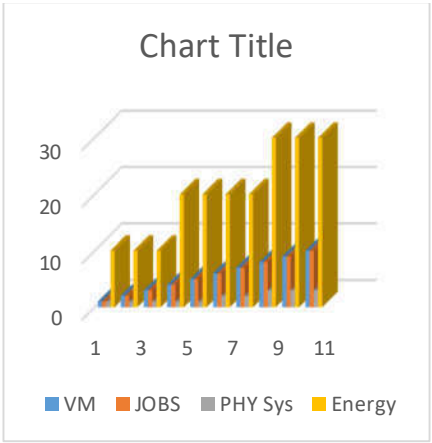
Graph:1 Energy Consumption with one PSystem



Graph:2 Energy Consumption with two PSystem



Graph:3 Energy Consumption with three PSystem



So in the above graph it is observed that if the virtual machines are used till the threshold capacity of the physical server is reached then the number of physical systems usage will be reduced there by reducing the consumption of energy. For the same number of virtual machines if one physical system is used then the graph:1 shows the energy consumption ,where as for same number of virtual machines two or three physical systems are used then graph:2 and graph:3 shows the consumption of energy.

Conclusion:

From the above algorithm and experiment work which is carried out the conslusion is that the number of physical systems to be used should be directly proportional to the number of virtual machines which can be launched on a physical system till the threshold capacity of the physical system is reached. Doing so we can straight away reduced lot of enegry consumption in the data center. Which will make the cloud computing as green computing and will also help the user to save lot on infrastructure utilisation.

References:

1. D. Saranya; Balancing Load Algorithms in CC.
A Review International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, Issue 7, 2015 July.

2 S. Sethi . Efficient Load Balancing in Cloud Computing using Fuzzy Logic, IOSR Journal of Engineering
(IOSRJEN) ISSN: 2250 -3021 vol. 2, pp. 65-71, 2012.

3 T. Desai. A Survey of Various Load Balancing Techniques and Challenges in Cloud Computing. International
Journal of Scientific & Technology Research, vol. 2, Nov 2013.

4 S. Rajoriya. Load Balancing Techniques in Cloud Computing: An Overview, International Journal of Science and
Research (IJSR), vol. 3, July 2014

5 Sharma S. et.al, “Performance Analysis of Load Balancing Algorithms,” World Academy of Science, Engineering and
Technology, 38, 2008.

6 Gross D. et.al, “No cooperative load balancing in distributed systems”, Elsevier, Journal of Parallel and Distributed
Computing, No. 65, pp. 1022-1034, 2005.

7 Nikravan M. “A Genetic Algorithm for Process Scheduling in Distributed Operating Systems Considering Load
Balancing”, Proceedings 21st European Conference on Modelling and Simulation (ECMS), 2007.

8 M. Amar “Service level agreement Driven Load Balancing for Web Applications in Cloud Computing Environment”,

9. Tian W, Xu M, Chen A. Open-source simulators for cloud computing: Comparative study and challenging issues. 2015:239–254.

10.Sunil Tekale Evaluation based load balancing
International Journal of Engineering & Technology.

11. W Tian Xu M, Y Zhong .
A kit for simulation/ modeling of real-time VM allocation in CC. IEEE Trans Autom Sci Eng. 2016

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

WEB INNOVATION WITH IOT IN SOCIAL ENVIRONMENT FOR SHARING EFFICIENT INFORMATION

1. **M.Kranthi,Assistant Professor,kranthimandhumulla@gmail.com**
2. **K.Y.Maheswari,Assistant Professor,maheswari.ky@gmail.com**
3. **K.Arun,Assistant Professor,karun014@gmail.com**

Abstract

IoT is a dream of interconnected regular items and self-setup remote systems of sensors. In IoT condition, it is expected to share data inescapably without client's mediation. We propose another interconnection method for productive data parttaking in IoT condition considering informal community. We introduce a technique and calculation which depends on not just the examination of the human's informal community yet in addition the thought of the gadget's sociality. At that point we depict a few situations and execute model framework utilizing the situations. A few examinations are led. From the trial assessment, we confirmed that our proposed method is useful in the effective collaboration between gadgets with no intercession of people.

Keywords: Pervasive figuring, Social systems, IoT, Information sharing

1. Introduction

There are logical confirmations that countless tied in an informal organization can give much more precise responses to complex issues than a solitary individual or a little gathering of people. The abuse of such a rule has been broadly examined in web related looks into. Some exploration have been proposed plans that utilization informal communities to seek web assets, to course activity, to choose successful arrangements for content circulation [1, 2]. With the development of web innovation, the idea of the interpersonal organization is increasingly connected in wide area of data innovation (IT). Individuals utilize an ever increasing number of electronic gadgets at home, in their auto or in the workplace, with the multiplication of keen associated gadgets [3].The Internet of Things (IoT) is a dream of interconnected ordinary articles and self-arrangement remote systems of sensors [4]. In this vision everything can be spoken to carefully and it will be conceivable to connect data and rationale to all things. The idea of IoT alludes to the innovation that empowers the correspondence between various gadgets, for example, advanced mobile phones, IP(internet convention) cameras, IP TV, associated autos or individual wellbeing gadgets without or with constrained client intercession [5].



Figure 1. The Concept of IoT

ISBN 978-81-939386-1-4

In the earth of IoT, the correspondence is utilized for robotized information transmission and estimation between mechanical or electronic gadgets. With the idea of IoT, the term data sharing is connected with cases which involve gadgets utilizing different inter connection procedures. Like portrayed in Figure 1, IoT applications in day by day activities can be found in different regions, for example, remote checking (e.g., sun oriented boards, business printer administrations), track and follow (e.g., rental bikes, armada following), office administration (e.g., lifts, vitality administration), distributing section (e.g., candy machines), or metering (e.g., utility meters). More application territories will emerge later on [6, 7].

In the IoT condition, everything between objects is black boxed. Accordingly, most clients are oblivious in regards to its inalienable highlights. Be that as it may, when clients need to share data in IoT condition, it is as yet required a troublesome manual process done by individuals themselves. To tackle this issue, we propose an effective unavoidable interconnection system to process and offer data in IoT condition.

For unavoidable interconnection in IoT condition, we think about the meeting of two ideas, the interpersonal organization and the IoT. So, in this paper, we portray another idea of the social IoT as an inescapable cooperation worldview for the data partaking in the Web of gadgets. Our interconnection system depends on the human's informal organization as well as the gadget's sociality. We consider the sociality of every gadget as gadget's believability for inescapably dependable data partaking in the earth of IoT. With a specific end goal to choose the sociality of the gadget, we think about the spatial separation between remote gadgets and the kind of focused gadgets (e.g. advanced mobile phones, tablet or wristband) as situational sociality. Additionally, we think about every gadget's history information (e.g. the kind of the most recent shared data or the profile rundown of the most recent interconnected gadget) as chronicled sociality. Through the assessment alongside the advancement of the model, we checked that our proposed strategy is useful in the proficient communication between gadgets with restricted mediation of people.

2. Related Works

IoT needs the innovation that empowers data sharing between various gadgets without or with restricted client mediation. With the improvement of IoT, non-human protests, for example, cell phones enter interpersonal organization and begin emulating social relations with individuals [8, 9].Therefore, it is critical to consider the collaboration of human-to-human as well as the interconnection of gadget to-gadget keeping in mind the end goal to share the data in the earth of IoT. As of late, there have been a significant number of research exercises that examine the possibilities of coordinating long range informal communication ideas into nature of IoT. Joining of the gadget to-gadget and the interpersonal organization is picking up energy. This is because of the developing requests for conveying numerous alluring ramifications into a future world populated by clever items saturating the regular day to day existence of individuals [10].

Plans in light of social connections for sharing data have been proposed to build up

larger amounts of trust and enhance the productivity and viability of security arrangements [11]. A few inquires about which exploits the interpersonal organization for information dispersion have been proposed [12, 13].Formo and Gårdman proposed the idea of a Social Web of Things as a psychological model and a collaboration worldview for the IoT. In their investigation, the human mental model of social relations is a substantially more extravagant and more instinctive method for comprehension the "net workedness" that is so basic in a web of things [7].The first thought of socialization between objects has been presented by Holmquist et al [14]. In their paper, the emphasis was on arrangements that empower savvy remote gadgets or for the most part remote sensors, to build up brief connections.

As of late, the possibility that the IoT and the informal communities are not separated from each different has started to show up in the writing. Ning et al. imagines the eventual fate of the Internet as being described by what they name Ubiquitous IoT engineering, which takes after the interpersonal organization based edge work demonstrate. That work gives a sagacious diagram of the normal IoT organize structure [15]. Notwithstanding, it doesn't go for abusing the attributes of the informal organizations into the earth of IoT. Kranz et al., portray ramifications of the mix between the IoT and the interpersonal organizations [16].And Dmitry and Manfred presented an approach for coordinating sensors and informal communities for social setting mindful information revelation [17]. These investigations, nonetheless, don't depict how the social connections ought to be set up amongst gadgets and does not propose any arrangement in regards to the required design and the innovation that is required for the inescapable interconnection of the gadget to-gadget without the mediation of people. Iván et al., introduced an answer for giving social administrations on the vehicle in view of IP mixed media subsystem and IoT abilities. In the examination, they built up a social administration named Drive and Share which offers pertinent data to vehicles in the earth of IoT [18]. In spite of the fact that the vehicular administrations have the basic element that the IoT condition in which the administration is connected, be that as it may, the examination did not completely think about the element. In light of this disregarding, the troublesome intercession of people is consistently required for the interconnection of the gadget to-gadget. As portrayed in above, there are numerous examinations about the mix of the interpersonal organization and the idea of IoT. In any case, the greater part of these investigations has not yet considered the strategy to restrain the intercession so as to effectively share data between gadgets. Hence, the investigation about the innovation that makes it accessible to inescapably share data between gadgets without the mediation of people is required in the period of IoT.

3. Pervasive IoT Interconnection Techniques for Information Sharing

In this paper, for inescapable interconnection of the gadget to-gadget, we think about the union of two ideas, the informal organization and the IoT. We depict another idea of the social IoT as an

ISBN 978-81-939386-1-4

unavoidable collaboration worldview for the data partaking in the Web of gadgets. Additionally, we propose an interconnection system which depends on the examination of the human's inclination as well as the thought of the gadget's sociality for the effective data partaking in IoT condition.

In our proposed procedure, we for the most part deal with two key focuses. Right off the bat, for the proficient and dependable data sharing, we think about interpersonal organization investigation, the sort of gadgets and the examination of gadget's sociality. The gadget's sociality is chosen in light of the spatial separation between gadgets. Also, for the customized human-driven data sharing, we concern the inclination of every client to discover a gadget with which more exact and customized data is shared. Calculation 1 is characterized for our proposed system. The calculation requires inputs which are profiles of the focused on client and the focused on gadget which are as of now utilized by the focused on client. What's more, two arrangements of other gadget's profile and other client's profile are required as contributions of Algorithm 1.Algorithm 1 is comprised of four stages.

Algorithm 1. *Pervasive IoT interconnection for information sharing*

Input:, target user u_{target}, targeted device d_{target} , DP is the set of N device's profile, UP is the set of N user's profile, the threshold of device sociality θ_{ds}

Output: The list of Top-Ndevices*final_candiates*

1.*first_filtered_candidates*←a memory space for the list of candidates filtered with the device's sociality
2. *second_filtered_candidates*←a memory space for the list of candidates filtered with the user's preference
3. **foreach**device,$dp_i \in DP$**do**
4. $sociality_i \leftarrow MachineSociality(d_{target}, dp_i)$
5. **if** $sociality_i \geq \theta_{ds}$**then**
6. *first_filtered_candidates*← mp_i
7. **for each** device,$dp_j \in$ *first_filtered_candidates***do**
8. **for each**$up_j \in UP$**do**
9.$similarity_j \leftarrow UserSimilarity(d_{target}, up_j)$
10. **if** $similarity_j \geq \theta_{ds}$**then**
11. **if**up_j's own device$mp_j \in$*first_filtered_candidates***then**
12. *second_filtered_candidates* ← mp_j
13. **for each**device, $dp_k \in$*second_filtered_candidates***do**
14. $credibility_k \leftarrow SocialMachineCredibility(d_{target}, dp_k)$
15. *final_candidates*←$RankingDevices(dp_k, credibility_k)$
16.**return***final_candidates*

The initial step is for characterizing the sociality of every gadget. The capacity Device Sociality ascertains the estimation of the gadget's sociality in light of the spatial separation. In Device Sociality, RSSI (Received Signal Strength Indication) is considered as the gadget's sociality. RSSI is a measure of the power level that a remote gadget, for example, WiFi, is getting from the radio foundation at a given area and time. Numerous examinations have considered RSSI for characterizing area of gadgets in the indoor condition [19]. Additionally, the broadly considered component of RSSI is that the higher the RSSI the better the quality and speed of the correspondence through the remote gadgets. On account of our examination, we utilize RSSI with a specific end goal to gauge the relationship of gadget to-gadget.

$$\theta_{ds} = \frac{\sum_{i=0}^{n} RSSI_i}{n} \times 0.1 \quad (-1 \leq \theta_{ds} \leq 1) \qquad (1)$$

In Algorithm 1, θds is the edge of the estimation of gadget's sociality. Condition (1) is a computation for θds and n is the quantity of all gadgets which are having a place with the arrangement of gadgets. Through the initial step, gadgets which are

not socially shut to the focused on gadget are sifted through. With the thought of the current every gadget's area as the sociality, we could build up the system which make it conceivable to dependably and unavoidably get a gadget interconnected to different gadgets in the IoT condition. The second step is for the thought of customized human-driven interconnection of gadget to-gadget. To choose whether a gadget can impart reasonably intriguing data to the focused on gadget, we think about the likeness between client's inclinations. In Algorithm 1, the capacity User Similarity restores the estimation of the closeness of the focused on client and another client upj. $similarityij$depicted in Equation (2) is the estimation of comparability of usersui and uj.And$similarityij$is figured with the estimation of Euclidian separation.

$$similarity_{ij} = \frac{(I_i - C_{ij}) \times (I_j - C_{ji})}{\sqrt{(I_i - C_{ij})^2 \times (I_j - C_{ji})^2}} \quad (2)$$

In Equation (2), Ii is the recurrence of data experienced by a client ui. Furthermore, cij is the recurrence of data regularly experienced by a client ui anduj. With the thought of client's closeness, we could choose which a gadget has the generally high opportunity to impart fascinating data to focused client's gadget. For instance, if the inclination of a client ui is moderately comparative with the inclination of a useruj,the devicedi being a client uj' claim may be considered as legitimate gadget to impart data to a client ui's gadget. Along these lines, with the second step in Algorithm 1, we could make it plausible to produce suitable interconnection of gadget to-gadget with the thought of the customized human-driven data sharing. The third step is for choosing whether the gadget is more qualification than different gadgets having a place with the set created through the initial step and the second step of the Algorithm 1. The fundamental capacity in the third step is Social Device Credibility. Social DeviceCrediblity computes the estimation of every gadget's social validity. Condition (3) is for the computation of gadget's social believability.

$$credibility_i = socialParticipation_i + deviceSimilarWeight_i. \quad (3)$$

We accept that the gadget's credibilityis firmly related with the principle factor which chooses the productivity of the unavoidable interconnection of gadget to-gadget. To appraise whether the gadget is more accreditation or not in the gadget to-gadget condition, we consider two highlights,$ocialParticipationi$and$deviceSimilarWeighti$.The initial segment of $crediblityi$ is the gadget's authentic sociality,$socialParticipationi$.The$socialParticipationi$defined in Equation (4) demonstrates how the devicedi has been partaken in data partaking in the earth of IoT.

$$socialParticipation_i = \frac{1-d}{n} + d \times \sum_{d_j \in DN(d_j)}^{n} \frac{socialParticipation_i}{Link(d_j)} \quad (4)$$

The $s\ ocialParticipationi$ in Equation (4) depends on the broadly utilized informal organization investigation, PageRank [20]. For the computation of $socialParticipationi$, we expect that the gadget which has higher recurrence of data sharing is more certification than different gadgets which have generally bring down recurrence of data sharing. For the procedures of the third step in Algorithm 1,

ISBN 978-81-939386-1-4

we utilize the history information of every gadget with a specific end goal to make sense of every gadget's recurrence of data sharing. At the end of the day, we could create the gadget interpersonal organization in view of the history information. For instance, if the way that the devicedi was recently interconnected with the devicedj is put away as history information, we could characterize the social connection between the combine of adevices di and dj. In view of the created gadget interpersonal organization, we measure how the gadget partakes in the data imparting to the gadget having a place with the informal community of gadgets. In Equation (4), n is the aggregate number of gadgets which are having a place with the informal organization of gadgets DN. The capacity $Linkmi$ restores the quantity of edges from the hub of the devicedi to other gadget hubs. d is the likelihood that thedevicedi which has been interconnected with the devicedj could interface with the other devicedk. The common component of IoT is versatility. In the earth of IoT, objects are always and powerfully changed in the area, channel of interconnection or the focused on protest with which the question is associated. Thus, we consider this portability highlight of IoT as the likelihood esteem d keeping in mind the end goal to calculate$socialParticipationi$ delineated in Equation (4). By and large, the same number of learns about the investigation on the connection structure utilize the estimation of dwith 0.85, we likewise see d as 0.85 in our examination. The second piece of $crediblityi$ is$deviceSimilarWeighti$. In the earth of the IoT, there are not just different kinds of gadgets from minor gadgets like the hub of electrocardiogram (ECG) to home apparatuses, for example, IP fridge or brilliant TV yet additionally various sort of data like media, messages or pictures. In this way, strategies for the unavoidable interconnection need to consider these assortments of nature of IoT keeping in mind the end goal to productively share data between gadgets. For instance, in light of the fact that the devicedi is computerized photograph outlines, the devicedicannot share data of which compose is video. With this case, the interconnection between the device$di$and a music playerdj isn't legitimate and the data sharing between these gadgets is useless. With our proposed Algorithm 1, we accept that for productive data media sharing, the sort of every gadget must be considered in the interconnection strategy. Thus, $deviceSimilarWeighti$ is for expanding the nature of offering data to the thought of the kind of every gadget. Condition (5) measures devicedi's$deviceSimilarWeight$.

$$deviceSimilarWeight_i = \frac{\sum_{d_i,d_j \in DP}^{n} ic_{ij}}{ic_i} \quad (5)$$

In Equation (5), ici is the recurrence of the interconnection from the deviceici to different gadgets. And$icij$ is the quantity of the interconnection from the focused on devicedi to different gadgets which have the sort being same with the devicedj. On the off chance that the devicedk has the regular interconnection with the gadget dj, device Similar Weight of the devicedi which has the sort indistinguishable to the kind of the device$dj$is moderately high. With the

ISBN 978-81-939386-1-4

estimation of device Similar Weight ascertained through Equation (5), we make it attainable to unavoidably and precisely share data between gadgets without the intercession of people.

4. Pervasive Media Sharing System Prototype

We build up the framework model in which our proposed strategy is connected. Through the framework model, we appraise the plausibility of our proposed system in the earth of IoT for inescapable media imparting to the restricted intercession of people.

4.1. Sample Scenarios

In this segment, we display test situations which are for the most part about the interconnection of the gadget to-deice for unavoidable data partaking in the earth of IoT. Atzori et al. described the number of meaningful scenarios in the IoT [21].Our sample scenarios are modified from scenarios of [21]. Figure 2 provides a sketch of our sample scenarios.



Figure 2. The Sketch of the Sample Scenarios

The example situations portrayed in Figure 2 is depicted in the accompanying.

① JB's house is situated in the square where is worked by cutting edge eco-accommodating standards. In the square, each house is outfitted with a shrewd meter and sensors keeping in mind the end goal to oversee and measure vitality utilization amid the entire day. By methods for the social IoT arrange, the shrewd meter can trade data on the vitality utilization with reference to distinguishing proof of the vitality suppliers that best match the house needs, recognizable proof of the family apparatuses. A light in any house changes the shading as indicated by the vitality sparing level acquired by its proprietor. Connections of the social IoT are abused in this situation.

② Jessie is a business agent that much of the time moves via auto around the city to meet her clients. Tragically, the movement has expanded in as of late making her driving increasingly dangerous. In any case, by exploiting the social IoT arrange, her auto can assemble data progressively about activity clog along conceivable courses and to pick the best way to get to the gathering in the planned time. Finding the fitting wellspring of intriguing data in the social IoT organize is simple for the auto by the inescapable IoT interconnection procedure proposed in our investigation.

③ Hannah has quite recently purchased another journal. Toward the starting, she experiences considerable difficulties to interface with some system supplies, for example, printers or faxes. By misusing the relationship of different gadgets having a place with the social IoT coordinate with our proposed interconnection method, Hannah's scratch pad can discover a mate of the gadget that has effectively tended to a similar design issues and fix the issues.

4.2. Implementations

We implement prototype systems using scenarios in Section 4.1, to verify the feasibility of our proposed technique. Figure 3 shows overall architecture of our prototype system.

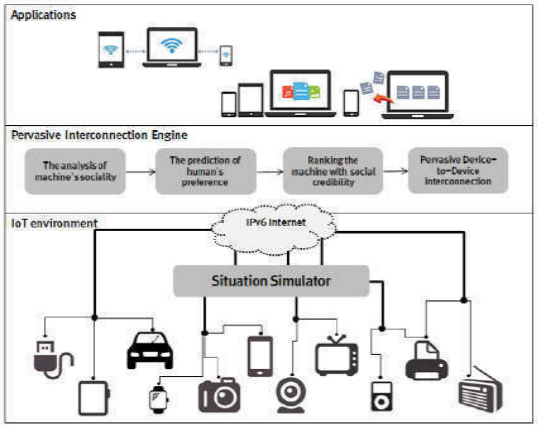


Figure 3. Overall Architecture of Prototype System

In the IoT condition appeared in the base of Figure 3, there are different gadgets which can associate the Internet. In our model framework, we executed a circumstance test system. The circumstance test system creates the circumstance in which our situations are appropriate. Inescapable IoT interconnecting motor portrayed amidst Figure 3 is the center of our model framework. The inescapable IoT interconnection motor is executed utilizing our procedure proposed in this paper. The unavoidable IoT interconnection motor is comprised of four stages which are depicted in Algorithm 1.

Table 1 depicts the usage condition with which the general model framework is created. With the execution condition portrayed in Table 1, we actualize a Web-based circumstance test system, inescapable interconnecting motor and applications which are adaptably material on two generally utilized versatile OS, iOS and Android.

Table 1. Implementation Environment

Components	Implementing tools
Situation simulator	MS SQL, IIS(Internet Information Services), .NET SDK, C#, Visual Studio 2015
Pervasive interconnecting engine	JDK 7, Java, Eclipse
Applications	JDK 7, ADT Marshmallow(Android Development Tools), iOS SDK, Java, Object C, Eclipse, Xcode

For the verification of the proposed technique, we especially more consider Scenario 3 than other scenarios. So, the measurement of device sociality in Figure 5(a) and the number of experimental results depicted in Chapter 5 are all based on Scenario 3.

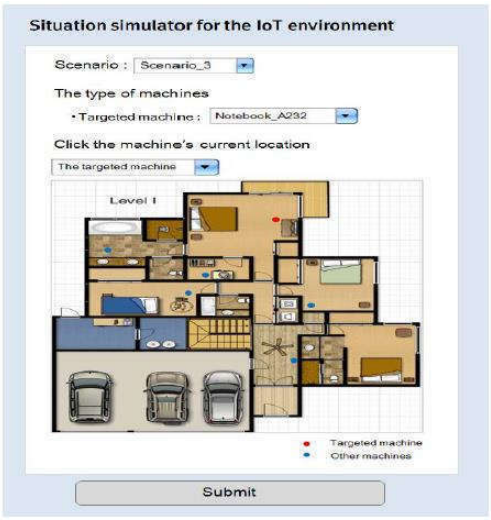
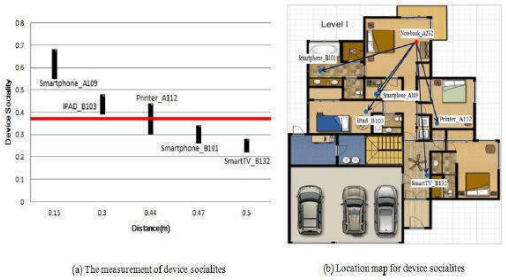


Figure 4. The Situation Simulator for the Scenario 3

Figure 4 demonstrates the circumstance test system. With it, we take the confirmation of our proposed procedure in view of the circumstance 3 depicted previously. Thus, we characterize the area of the focused on scratch pad and different gadgets with the circumstance test system.

ISBN 978-81-939386-1-4



(a) The measurement of device socialites (b) Location map for device socialites

Figure 5. The Analysis of Device Socialites

Figure 5(a) shows the measurement of each device's sociality based on the location of the device, Notebook_A232 depicted as the red circle in Figure 5(b). Like shown in Figure 5(a), the device sociality mainly depends on the distance from the targeted device.



(a) The existing application (b) The proposed application

Figure 6. The Comparison between Two Applications

Likewise, as we characterized the condition for the gadget sociality with estimations of gadget's RSSI, the higher estimation of the gadget sociality implies that it is conceivable to impart data between gadgets to better quality and quicker speed. The flat line implies the estimation of gadget's sociality depicted in Equation (1). Thus, on account of the estimation appeared in Figure 5(a), three gadgets, Smartphone_A109, iPAD_B103 and Printer_A112 are just considered as applicant gadgets for the interconnection to the focused on gadget, Notebook_A232. Figure 6 demonstrates the examination between the use of the current gadget recovery and the application in which our proposed interconnection system is connected. As appeared in Figure 6(a), the current application arbitrarily recovers the rundown of gadgets with signals got from different gadgets close to a focused on gadget. Then again, the application executed with our inescapable interconnection strategy gives the rundown of gadgets which are arranged by every gadget's noteworthiness. In the rundown, the higher a gadget is put in the rundown, the more conceivable the gadget may be interconnected with the focused on gadget. For instance, in Figure 6(b), the gadget, Printer_A112has the most astounding probability to interconnect with the focused on gadget, Notebook_A232 among different gadgets recorded in Figure 6(b). Likewise, while the current application records all gadgets close to the focused on gadget, Notebook_A232, the application actualized with our unavoidable interconnection technique retrieves just three gadgets. These tree gadgets could be viewed as more dependable, customized and certification than different gadgets which are sifted through our proposed procedure.

5. Evaluation

To proficiently confirm the execution of the model actualized with the proposed strategy, we directed the near trial on the two models named individually as Proposed and Existing. While the model Existing just records gadgets almost a focused on gadget, the model Proposed recovers the rundown of gadgets arranged with our proposed strategy. After the rundown of gadgets is recovered, the models existing haphazardly select a

ISBN 978-81-939386-1-4

gadget to interface with the focused on gadget. Something else, the model Proposed select the gadget which is situated in the highest point of rundown. At that point, two models individually make an impression on the chose gadget for sharing data between the chosen gadget and the focused on gadget. In the examination, likewise, we consider two variables of the data sharing between gadgets. The one is the productivity, and the other is the exactness for the data partaking in the earth of IoT. Initially, for the confirmation of the productivity of the data sharing, we assessed the execution of two models regarding the time required to share data between gadgets. Figure 7 demonstrates the consequence of the similar analysis on the productivity for the data sharing.

In Figure 7, the x-pivot is comprised of three focused on gadgets which are individually portrayed in three situations, Scenario 1, Scenario 2, and Scenario 3. Prior to the trial, we expected that the required time portrayed in Figure 7 as the y-hub, is add up to time required to discover suitable gadgets, with a specific end goal to make association amongst gadgets and forward data from a gadget to the next gadget. We gauges the required time of the every situation. Through the aftereffect of the examination for the productivity, we realize that the model, Proposed requires less time to share data between the specific gadget and the focused on gadget than the model, Existing. Since the model, Existing arbitrarily chooses a gadget having a place in the rundown of recovered gadgets, if the endeavor to make the association from the chose gadget to the focused on gadget is fizzle, the model retries the way toward discovering gadgets, choosing a gadget and making an impression on a chose gadget devouring increasingly time to share data. Then again, as the model Proposed just chooses the gadget situated in best of the rundown which is recovered with our proposed system, it is exceedingly conceivable to make the interconnection between the focused on gadget and the chose gadget at once without exercise in futility.

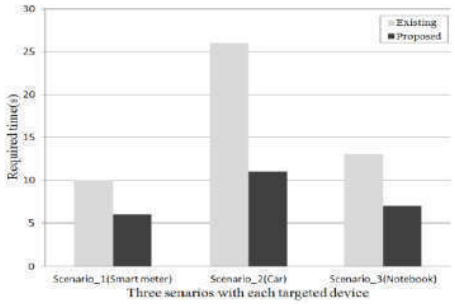


Figure 7. The Result of the Comparative Experiment on the Efficiency

At long last, to check the exactness of the data sharing, we assessed the execution of two models regarding the blunder rate happening during the time spent the sharing data. In the earth of IoT, while there may be different sorts of blunders, we just consider the quantity of mistake writes in our analysis. As such, we accept that every situation portrayed above has an appropriate gadget to impart data to the focused on gadget. For instance, on account of Scenario 2, in light of the fact that the focused on gadget is an auto and the client of the auto needs to share data about movement clog along conceivable courses, the association between the auto and a printer isn't precise. Along these lines, this case is viewed as the one of a blunder in data partaking in our test. Figure 8 demonstrates the aftereffect of our similar trial on the exactness of two models.

ISBN 978-81-939386-1-4

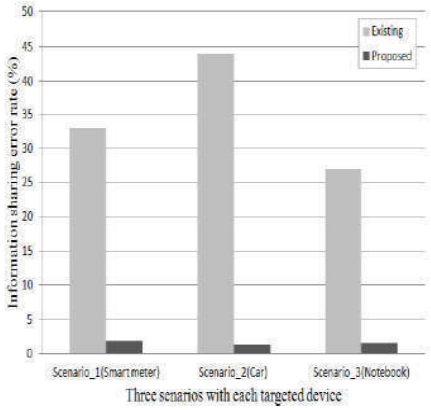


Figure 8.The Result of the Comparative Experiment on the Accuracy

In Figure 8, the x-pivot is comprised of three focused on gadgets which are separately depicted in three situations, Scenario 1, Scenario 2, and Scenario 3. Through the aftereffect of the trial for the precision, we realize that the model, Proposed has bring down mistake rates in all situations than the model, Existing. Since the model, Existing arbitrarily chooses a gadget having a place in the rundown of recovered gadgets, the odds of disappointment are more higher than the disappointment shot of the model, Proposed. Something else, the model, Proposed just chooses the gadget which is situated in the highest point of the recovered rundown. As the chose gadget is considered as the most legitimate gadget among different gadgets in the rundown by the Algorithm 1, the association and data sharing between the focused on gadget and the chose gadgets is for the most part fruitful.

6. Conclusion

In IoT condition, it is expected to share data unavoidably without client's mediation. In this paper, for inescapable interconnection in IoT condition, we think about the meeting of two ideas, the interpersonal organization and the IoT. Right off the bat, for the proficient and solid data sharing, we think about interpersonal organization examination, the kind of gadgets and the investigation of gadget's sociality. Furthermore, for the customized human-driven data sharing, we concern the inclination of every client keeping in mind the end goal to discover a gadget with which more precise and customized data is shared. This paper depicted our proposed technique and calculation. To assess the practicality of our proposed method, we executed model framework utilizing our situations. Likewise, we directed the similar trial on the model framework. Through the test, we could confirm that our proposed system make it conceivable to unavoidably share data between gadgets without the mediation of the human with the method for more exactness and more effectiveness.

References

[1] S. Marti, P.Gansna and H. G. Molina, “SPROUT: P2P routing with social networks”, Proceedings of Current Trends in Database Technology-EDBT Workshops (2004).

[2] A. Fast, J. David and B. N. Levine, “Creating social networks to improve peer-to-peer networking”, Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM (2005).

[3] J. Woo, “Information Retrieval Architecture for Heterogeneous Big Data on Situation Awareness”, International Journal of Advanced Science and Technology, (2013), pp. 113-122.

[4] G. Wu, S. Talwar, K. Johnsson, N. Himayat and K. D. Johnson, “IOT: From mobile to embedded internet”,

Communications Magazine IEEE, no. 4, (2011), pp. 36-43.
[5] M. J. Booysen, J. S. Gilmore, S. Zeadally and G. J. Van Rooyen, "Machine-to-machine (m2m) communications in vehicular networks", , KSII Transactions on Internet and Information Systems (TIIS), no. 2, (2012), pp. 529-546.
[6] V. B. Misic and J. Misic, "Machine-to-Machine Communications: Architectures, Technology, Standards, and Applications", CRC Press, (2014).
[7] J. Formo, L. Jarmo and M. Gårdman, "Internet of things marries social media", Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, ACM, (2011).
[8] R. J. Robles and T. Kim, "Application, Systems and Methods in Smart Home Technology: A Review", International Journal of Advanced Science and Technology, vol. 15, (2010), pp. 37-47.
[9] R. Sandip, B. Rahesh Bose and S. Debabrata, "A Fog-Based DSS Model for Driving Rule Violation Monitoring Framework on the Internet of Things", International Journal of Advanced Science and Technology, vol. 82, (2015), pp.23-32.
[10] L. Atzori, I. Antonio, M. Giacomo and M. Nitti, "The social internet of things (siot)–when social networks meet the internet of things: Concept, architecture and network characterization", Computer Networks, vol. 56, no. 16, (2012), pp. 3594-3608.
[11] H. Yu, M. Kaminsky, P. B. Gibbons and A. Flaxman, "Sybilguard: defending against sybil attacks via social networks", ACM SIGCOMM Computer Communication Review, vol. 36, no. 4 (2006), pp. 267-278.
[12] P. Costa, C. Mascolo, M. Musolesi and G. P. Picco, "Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks", Selected Areas in Communications, IEEE Journal, vol. 26, no. 5, (2008), pp. 748-760.
[13] L. Ding, S. Peng and B. Liu, "The clustering of internet, internet of things and social network", Proceeding of Knowledge Acquisition and Modeling (KAM), 3rd International Symposium, IEEE, (2010).
[14] L. E. Holmquist, M. Friedemann and S. Bernt, P. Alahuhta, M. Beigl and H. W. Gellersen, "Smart-its friends: A technique for users to easily establish connections between smart artefacts", Proceeding of Ubicomp: Ubiquitous Computing, Springer Berlin Heidelberg, (2001).
[15] H. Ning and Z. Wang, "Future Internet of things architecture: like mankind neural system or social organization framework?", Communications Letters, IEEE, vol. 15, no. 4, (2011), pp. 461-463.
[16] M. Kranz, L. Roalter and F. Michahelles, "Things that twitter: social networks and the internet of things", Proceeding of What can the Internet of Things do for the Citizen (CIoT) Workshop at The Eighth International Conference on Pervasive Computing, (2010).
[17] M. Baqer and A. Kamal, "S-sensors: Integrating physical world inputs with social networks using wireless sensor networks", Proceeding of Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 5th International Conference, IEEE, (2009).

[18] I. Lequerica, M. G. Longaron and P. M. Ruiz, "Drive and share: efficient provisioning of social networks in vehicular scenarios", Communications Magazine, IEEE, vol. 48, no. 11, (2010), pp. 90-97.
[19] K. Kaemarungsi, “Distribution of WLAN received signal strength indication for indoor location determination”, Proceeding of Wireless Pervasive Computing, 2006 1st International Symposium, IEEE, (2006).
[20] A. Mislove, M. Massimiliano, K. P. Gummadi, P. Druschel and B. Bhattacharjee, “Measurement and analysis of online social networks”, Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, ACM, (2007).
[21] L. Atzori, I. Antonio, M. Giacomo and M. Nitti, “The social internet of things (siot)–when social networks meet the internet of things: Concept, architecture and network characterization”, Computer Networks, vol. 56, no. 16, (2012), pp. 3594-3608.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A SURVEY ON AI DRIVEN CLOUD COMPUTING-FUTURE TRENDS

B.Anitha K.Divya Bharathi J.Srinivas

Asst.Professor Asst.Professor Asst.Professor

anithabejugama@gmail.com kotadivya6@gmail.com jsrinivas@mjcollege.ac.in

Abstract: This paper demonstrates the role of Artificial Intelligence in the field of Cloud Computing. It surveys the basic applications of Artificial Intelligence in Computer Science. The paper explains how the power of AI can be harnessed to optimize cloud computing. It also discusses about the advantages of AI driven Cloud Computing. Finally the paper deliberates on the future trends of AI driven Cloud Computing.

Keywords:- Artificial Intelligence, Cloud Computing, Machine Learning,

I. Introduction

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [1]. Cloud computing has evolved as a popular answer to provide economic and simple access to externalized IT (Information Technology) facility [2]. With the beginning of Infrastructure-as-a-Service, huge-scale Software systems like community networking sites and e-commerce software are being developed in order to harness the power of cloud [3].An end user can unilaterally provision computing ability, such as server time and network storage, as needed involuntarily without requiring human interface with each service's provider. Access to your information from anywhere at any time is the specialty of hosted environments. You don't need to be in the same physical location as the hardware that stores your data. The Cloud provider houses the hardware and software necessary to run your applications.[4]

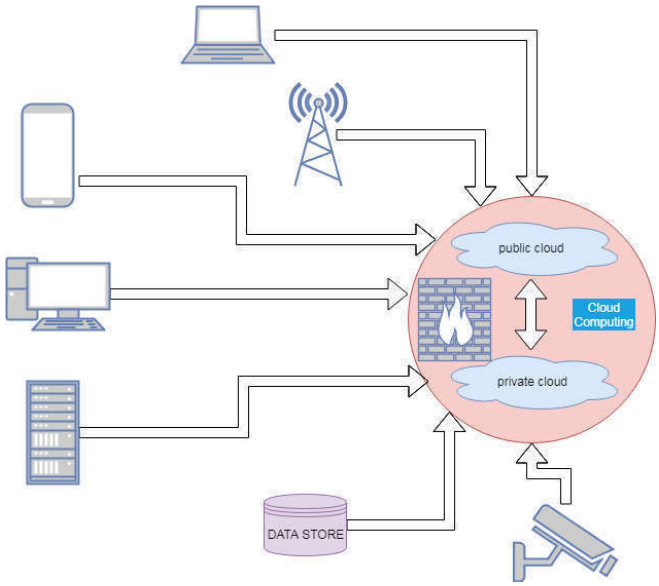


Fig 1. Cloud Computing Frame work

Artificial Intelligence, popularly referred to as AI, refers to the simulated intelligence in machines[5]. The term refers to the end

result of endowing machines with the rational prowess peculiar to humans, the ability to reason, learn from the past, discover meaning, or generalize. It is hinged on the principles that human intelligence can be defined in such precise terms that a machine can mimic it. These machines are hence programmed to "think" just like a human would and imitate the actions and reactions of humans to certain circumstances. The cloud and AI merge perfectly in various ways and according to scholars; AI might just be the technology to revolutionize cloud computing solutions. AI as a service improves the existing cloud computing results and engenders new paths to expansion. Figure 2 depicts a survey done by us on how AI will impact human life.

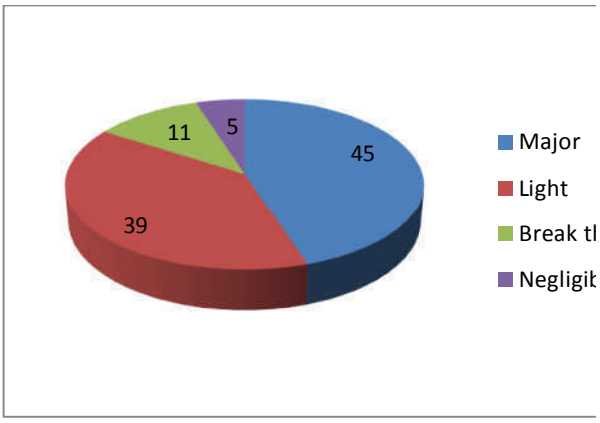


Fig 2. Impact of AI (5 years from now)

Cloud computing services have morphed from plans such as Google App Engine and Azure to Infrastructure which involves the provision of machines for computing and storage. In totaling to this, cloud providers also offer data podium services which cross the dissimilar existing databases. This series of expansion points in the direction of the growth of Artificial Intelligence and Cloud Computing. Almost all tech savvy people these days know about the potential of cloud computing technology and how cloud platform has already artificial businesses by effectively storing data and matching the existing workloads. Because cloud computing is a break though technology, many organizations need time to think and analyze how cloud will evolve with time. Figure 3 depicts the overall spending in $ billions by various IT companies from 2015 to 2020. [Source:- IDC 2016]

Cloud technology saw key changes when mobile phones started substituting computers and the Internet of Things (IoT) platform came into picture. The subsequent point now is to see how Artificial Intelligence (AI) improvises cloud taking into account the way cloud

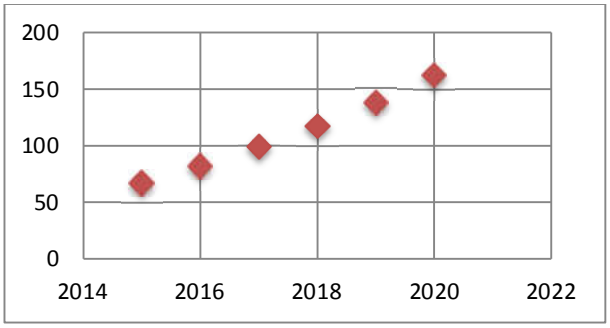


Fig 3. Rapid Growth of Cloud Computing

technology has improvised AI development. Figure 4 talks about the various disciplines where AI can be useful.

ISBN 978-81-939386-1-4

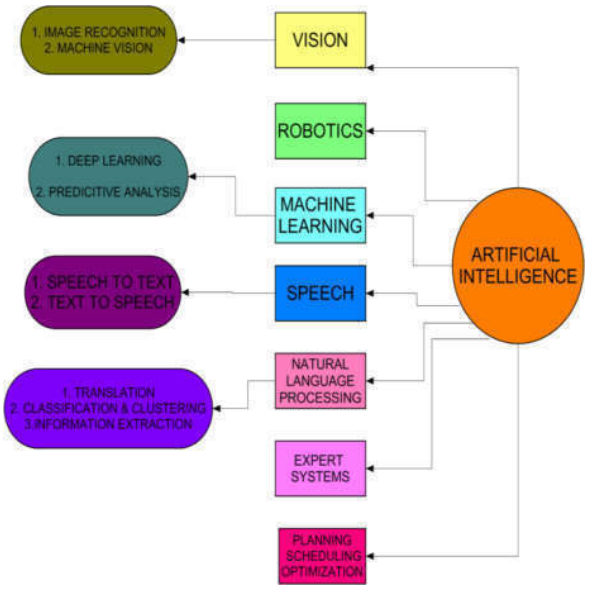


Fig.4. AI & related fields

Currently AI is serving in areas like Vision, Robotics, Machine Learning, Speech, NLP, Expert Systems, Planning, scheduling and optimization. All these areas require large processing power and storage capacity which can be provided by cloud only.

II. Artificial Intelligence enabled by cloud

About 90% of untimely cloud adopters assert that cloud computing will play a foremost role in their Artificial Intelligence applications in future. And more than 55% of users chose cloud-based services and are leveraging SaaS and PaaS to carry out and deploy AI-infused cloud fallout. The early adopters have also shared their experiences declaring that enabling cloud technology will play a significant role in AI adoption. Hence, we can say that persistent AI is supported by persistent cloud. The cloud technology is a well-established technology trend that is majorly ruled by IT companies like ESDS, Google, Amazon, etc. AI definitely carries unique features that can influence various next generations of cloud computing platforms. AI requires support for brand new programming paradigms and also needs a new computing infrastructure. So, we can expect AI capabilities to be held and integrated by the cloud as a standard element of its infrastructures. Also we can anticipate seeing the initiation of a new generation of cloud platforms powered by AI. Maybe, we are entering the epoch of the AI-first cloud. A majority cloud of stake holders strongly believe that the fusion of cloud computing services and AI technology will bring a significant transformation in the technology industry. Public cloud providers keep on investing in the expansion of AI and this will continue to draw the right set of customers to this technology. Even though the technology is still in its early stage, the development to come is inevitable and we can anticipate phenomenal advancements in the future. Cloud computing utilizing AI is certainly not a radical or progressive change. In numerous regards, it's a transformative one. For some organizations, it has been a flawless integration from existing frameworks, with AI investment gathering pace rapidly.

III. AI Infrastructure for Cloud Computing

There are two prominent ways in which AI can steer Cloud Computing

1. Cloud Machine Learning (ML) Platforms.
2. Technologies that support AI platform for business that is hosted on Cloud.

Cloud Machine Learning (ML) Platforms:

ISBN 978-81-939386-1-4

We can spawn Machine Learning (ML) models when a bulky set of data is applied to certain algorithms, and it becomes significant to leverage the cloud for this. The models are able to be trained from the different patterns which are gleaned from the accessible data. As we supply more data for this model, the prediction gets superior and the accuracy is better. For instance, for ML models which recognize tumors, thousands of radiology reports are used to train the system. This prototype can be used by any industry since it can be customized based on the project needs. The data is the requisite input and this comes in different forms - raw data, unstructured data, etc. Because of the advanced computation techniques which require a blend of CPUs and GPUs, cloud providers now offer virtual machines with incredibly potent GPUs. Also, machine learning tasks are now being programmed by means of services which comprise batch processing, serverless computing, and orchestration of containers. IaaS also aids in conducting predictive analytics. Modern day technologies like AWS ML, Azure ML and the forthcoming Google Cloud ML use a precise technology that is held accountable for powering the conception of Machine Learning models. But excepting Google Cloud ML that leverages TensorFlow can be hard since a huge quantity of cloud ML technologies won't authorize implementation of AI programs coded in conventional AI.

Technologies that support AI platform:

Even without creating a unique ML model, it is likely to enjoy services which parallel that provided by the AI systems. For instance, text analytics, speech, vision, and machine language translation are available to developers. They can merely integrate this into their development projects. Although these services are common and are not tailored to specific uses, cloud computing vendors are taking steps to guarantee that this is frequently improved. Cognitive computing is a model which permits users to provide their adapted data which can be taught to deliver well-defined services. This way, the trouble of finding the suitable algorithm or the correct training model is purged. Technologies that support AI platform for industry like IBM Watson, Google Cloud Vision, Microsoft Cognitive Services or Natural Language application programming interfaces permit abstract complex AI capabilities via simple API calls. Using this you can slot in AI capabilities without investing in complicated AI infra. The AI technologies are bound to progress with time and cloud platforms will toggle from the level of basic support for AI capabilities to a much flexible model where AI programs are as extensively supported as web and databases function these days.

IV. Benefits of Leveraging AI and Cloud Computing

A. Cost-Effectiveness

By being available through the internet, the cloud application development removes the need for expenses on-site hardware and software procurement and setup. It also eliminates the need for on-site data centers and the operating cost that arrive with it – IT professionals to manage the centers, servers and round the clock power to power and cool the servers.

B. Increased Productivity

Unlike a hard drive or local storage device which needs a lot of IT management chores – hardware setup, software patching, racking and stacking – cloud computing is all internet based and as such has no need for this. This gives room for the IT team to spotlight on achieving other business goals using AI.

C. Reliability

With a hard drive or physically accessible infrastructures, the risk of damage is heightened. One faces the risk of the crash, lost files, backup failure and so much more. However, cloud computing solutions ensure business continuity, faster and easier disaster recovery and easier data backup.

D. Availability of Advanced Infrastructure

AI applications are normally high performance when on servers with multiple and extremely fast Graphics Processing Units (GPUs). These systems are nevertheless extremely expensive and high-priced for many organizations. AI as a service in cloud application development becomes accessible to these organizations at a more reasonable price.

V. Conclusion

Our paper introduced the concepts of Cloud Computing & Artificial intelligence. It discussed the various areas where AI powered Cloud can power up reasonable solutions. Finally our paper draws out the following Conclusion “Throughout the next couple of years, we can hope to see the business keeps on blasting, with AI driving cloud computing higher than ever, while the cloud business conveys the advantages of AI to the mainstream”.

REFERENCES

1. "NIST Cloud Computing Definition", NIST SP 800- 145].
2. CLOUD COMPUTING BASICS J.SRINIVAS et.al International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 5, July 2012.
3. J.SRINIVAS et.al Exploiting Geo Distributed Datacenters of a Cloud for Load Balancing 2015 IEEE International Advance Computing Conference (IACC).
4. J.SRINIVAS et.al A novel approach for data storage security in a hosted environment International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.3, No.1, February 2013.
5. https://www.goodfirms.co/blog/role-of-ai-in-cloud-computing.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Reorganization of DFA based prototype Matchers

Ch Vengaiah
Assistant Professor, MRCE,
Department of CSE
Vengaiah19@gmail.com

K Praveen Reddy
Assistant Professor,MRCE,
Department of CSE
Kpreddy534@gmail.com

M.Aharon
Assistant Professor,MRCE,
Department of CSE
aharon.mattakoyya@gmail.com

ABSTRACT

This paper presents an efficient algorithm for finding matches to a given regular expression in given text using optimization of DFA. To match a regular expression of length n, a serial machine requires 0(2^n) memory and takes 0(1) time per text character. The proposed approach requires only 0(n^2) space and still process a text character in 0(1) time (one clock cycle).The improvement is due to the optimization of DFA that means without converting it into the NFA, directly convert into the DFA. Finite Automaton (DFA) used to perform the matching. Furthermore, the paper presents a simple, fast algorithm that quickly constructs the DFA for the given regular expression.

Keywords

DFA, NFA , Regular Expression

I. INTRODUCTION

A regular expression, often called a pattern, is an expression used to specify a set of strings required for a particular purpose. Regular expressions are widely supported in programming languages, text processing programs (particular lexers), advanced text editors, and some other programs. A regular expression processor translates a regular expression into a nondeterministic finite automaton (NFA), which is then made deterministic and run on the target text string to recognize substrings that match the regular expression. The pattern sequence itself is an expression that is a statement in a language designed specifically to represent prescribed targets in the most concise and flexible way to direct the automation of text processing of general text files, specific textual forms, or of random input strings.This paper presents an efficient algorithm for finding matches to a given regular expression in given text using optimization of DFA.

II. BACKGROUND

followpos(i) : is the set of positions which can follow the position i in the strings generated by the augmented regular expression.

followpos is just defined for leaves,it is not defined for inner nodes.

Computing Followpos:

A position of a regular expression can follow another position in two ways:

Rule 1: if n is a cat-nodec1c2

For every position in lastpos(c1) all positions in firstpos(c2) are in followpos(i).

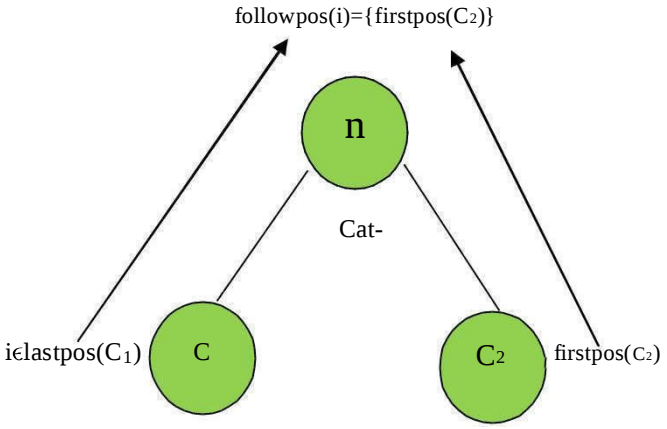


Figure 1: Cat -Node

Rule 2:if n is a star-node

If i is a position in lastpos(n) then all positions in firstpos(n) are in followpos(i).

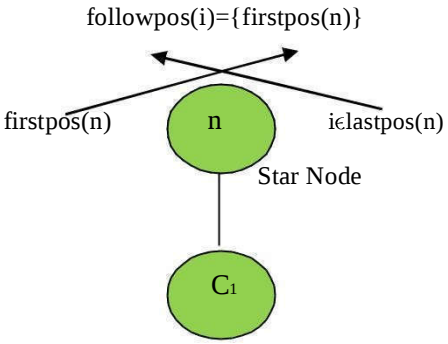


Figure 2: Star-Node

Evaluation of firstpos, lastpos, nullable

To evaluate followpos, three more functions are to be definedfor the nodes (not just for leaves) of the syntax tree.

nullable(n) true if the empty string is a member of strings generated by the sub-expression rooted by n false otherwise

firstpos(n)the set of the positions of the first symbols of strings generated by the sub-expression rooted at n.

lastpos(n) the set of the positions of the last symbols of strings generated by the sub-expression rooted at n.

Table:1Evaluation of firstpos, lastpos, nullable

Node n	Nullable(n)	Firstpos(n)	Lastpos(n)
A leaf labeled ε	True	Ø	Ø
A leaf with position i	False	{i}	{i}
An or-node n=c1|c2	Nullable(c1) or nullable(c2)	Firstpos(c1) U firstpos(c2)	Lastpos(c1) U lastpos(c2)
A cat-noden=c1c2	Nullable(c1) and nullable(c2)	If (nullable(c1)) Firstpos(c1)U firstpos(c2) Else firstpos(c1)	If(nullable(c2)) Lastpos(c2) U lastpos(c1) Elselastpos(c2)
A star-noden=c1*	True	Firstpos(c1)	Lastpos(c1)

Need Of Pattern Matching

Pattern matching is the process of checking a perceived sequence of string for the presence of the constituents of somepattern. In contrast to pattern recognition, the match usually has to be exact. The patterns generally have the form sequencesof pattern matching include outputting the locations of a pattern within a string sequence, to output some component of thematched pattern, and to substitute the matching pattern with some other string sequence (i.e., search and replace). Patternmatching concept is used in many applications Following figure shows the different applications.

In pattern matching I focused on the regular expression amongst others application.Pattern matching will help to find right and appropriate result. Hence I proposed Streamlining of DFA Based Pattern Matchers algorithm to match the Pattern.



Figure 3: Applications of Pattern Matching

III. PROPOSED ALGORITHM

Algorithm to make RE to DFA:

1. Create syntax tree with augmented rerular expression.
2. Assign number to each terminal from left to right.If there is any epsilon as a terminal,do not assign any number.
3. Calculate the functions: followpos, firstpos, lastpos, nullable.
4. Put S(start node)= firstpos (root of augmented syntax tree) of DFA as an unmarked state.
5. Mark S and make it start state.
6. Now we give the input to the start state which is already marked.
7. Input is provided according to numbers which are in given states.
8. If terminals are defined in more than one position then find the followpos of those terminals on different positions and do the union.
9. Now we find out the next state which is move(S,a)
   = S'(next state which can also be final state).
10. If the next state is new state then go to step[7].
11. If the next state is repeated state and no new state is found then stop process.
12. To make DFA according to move fuction.
13. The accepting states of DFA are all states containing the position of #.

Example: Augmented Regular Expression

(a | b) * a #

red – firstpos, blue – lastpos

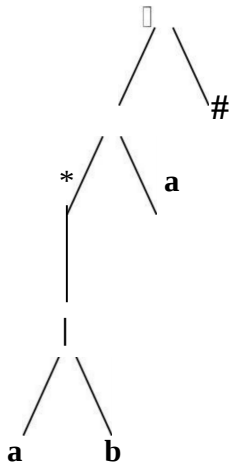


Figure 4: Syntax tree with augmented regular expression

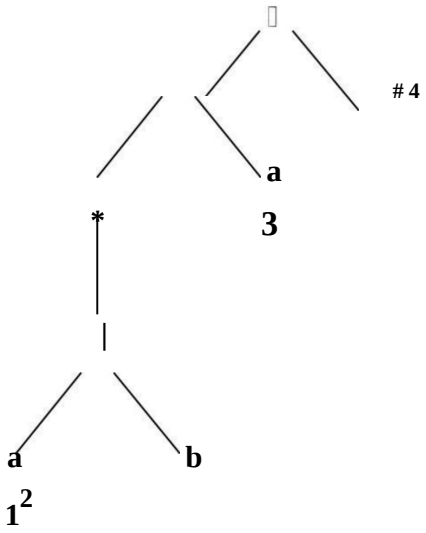


Figure 5: Assign number to each terminal from left to right

Calculate the functions: followpos, firstpos, lastpos, nullable

Calculate firstpos(i) and lastpos(i) for leaf i:

For i=1,

firstpos(1): {1}

lastpos (1): {1}

For i=2,

firstpos(2): {2}

lastpos (2): {2}

For i=3,

firstpos(3): {3}

lastpos (3): {3}

For i=4,

firstpos(4): {4}

lastpos (4): {4}

Calculate firstpos(i) and lastpos(i) for or, cat and star node:

For or node(|):

firstpos(|): {1,2}

lastpos(|): {1,2}

For star node(*):

firstpos(*): {1,2}

lastpos(*): {1,2}

For cat node(.):

For leaf a:

firstpos(.): {1,2,3}

lastpos(.): {3}

For leaf #:

firstpos(.): {1,2,3}

lastpos(.): {4}

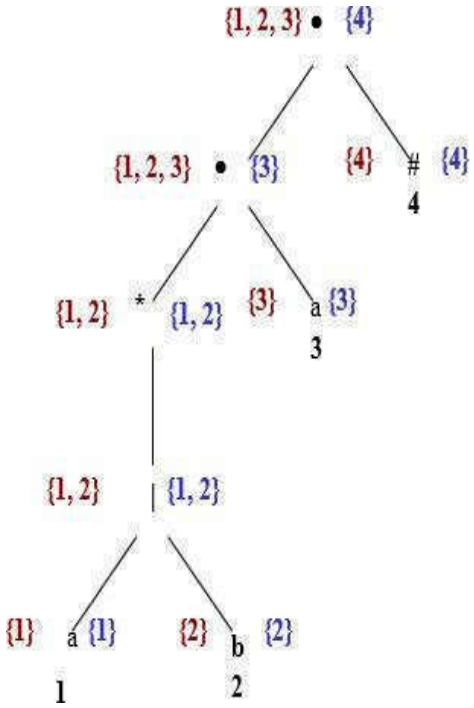


Figure 6: Augmented ST with computed values

followpos(1) = {1,2,3}

followpos(2) = {1,2,3}

followpos(3) = {4}

followpos(4) = {}

S_1=firstpos(root)={1,2,3}

mark S_1

a: followpos(1) ∪ followpos(3)={1,2,3,4}= S_2, move(S_1,a)=S_2

b: followpos(2)={1,2,3}=S_1, move(S_1,b)=S_1

mark S_2

a: followpos(1) ∪ followpos(3)={1,2,3,4}= S_2, move(S_2,a)=S_2

b: followpos(2)={1,2,3}=S_1move(S_2,b)=S_1

start state: S_1

accepting states: {S_2}

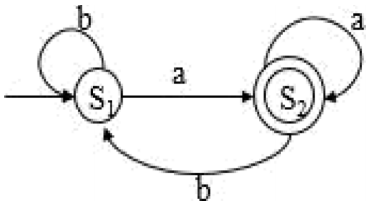


Figure 7: Deterministic finite automata

IV. CONCLUSION

In this paper, we proposed an efficient regular expression algorithm to match the regular expression without converting into NFA.Which results the reduced area, better performance, less number of resources.To match a regular expression of length n, a serial machine requires 0(2^n) memory and takes 0(1) time per text character. The proposed algorithmrequires only 0(n^2) space and still process a text character in 0(1) time (one clock cycle) .The improvement is due to the optimization of DFA that means without converting it into the NFA, directly convert into the DFA. The future scope of the work is to improve the efficiency of the finite state machine.

V. REFERENCES

[1] J. E. Hopcroft and J. D. Ullman, Introduction to automata theory,languages, and computation," *Addison Wesley*, 1979.

[2] Y. Sun, H. Liu, V. Valgenti, and M. S. Kim, "Hybrid regular expression matching for deep packet inspection on multi-core architecture," in *Proceedings of the 19th International onference on Computer Communications and Networks, ICCCN'10*, Aug. 2010, pp. 1–7.

[3] P.Jayaprabha and Rm. Somasundaram ,"Content Based Image Retrieval Methods Using Graphical Image Retrieval Algorithm (GIRA)"., Computer Science And Application, Vol. 1, No. 1, January, 2012.

[4]1997,http://www.igm.univmlv.fr/~lecroq/string/index.html C. Charras and T. Lecroq: Exact String MatchingAlgorithms. Univ. de Rouen.

[5] Srikanthan, Sharanyan ,"Implementing the dynamic time warping algorithm in multithreaded environments for realtime and unsupervised pattern discovery"., Computer and Communication Technology (ICCCT), 2011 2nd International Conference on 394 – 398, 15-17 Sept. 2011.

[6] Stan Salvador & Philip Chan,"Fast DTW: Toward Accurate Dynamic Time Warping in Linear Time and Space".KDDWorkshop on Mining Temporal and Sequential Data, pp. 70-80, 2004. (http://cs.fit.edu/~pkc/papers/tdm04.pdf).

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Protected and Flexible Multi-Keyword Search model over Encoded Cloud Data

Syed mohammed shafi,Assistant Professor in CSE Dept,MRCE
syd.shafi16@gmail.com

Abstract: Cloud computing is modern technology as a new computing model in number of business domains. Large numbers of large scale departments are starting to shift the data on to the cloud environment. With the benefit of storage as a service many enterprises are moving their valuable data to the cloud, since it costs less, easily scalable and can be accessed from anywhere any time. Improved dynamic multi-keyword ranking search scheme with top key via encrypted cloud data that simultaneously supports dynamic update operations as deleting and inserting documents. Greedy depth first search algorithm is provided for efficiency multi keywords on place and index structure. Cryptography is one of the establishing trust models. Searchable security is a cryptographic method to provide security. In number of researchers have been working on developing privacy and efficient searchable encryptiontypes. We take new effective cryptographic techniques based on data structures like CRSA and B-Tree to enhance the level of privacy. We propose new multi-keyword search query over encrypted cloud information in retrieving top k scored documents. The vector space model and TFIDF model are used to build index and query generation. This paper focuses on multi keyword search based on ranking over an encrypted cloud data. The search uses the feature of similarity and inner product similarity matching. We propose to support the top-k Multi-full-text search for security and performance analysis show that the proposed model guarantees a high safety and practicality and dynamic update operations, such as deleting and adding documents. The experimental results show that the overhead in computation and communication is low.

Index Terms: *Advanced Symmetric Encryption Certified Authority, Cloud data, -Multi keyword Retrieval, Cloud data, Data security, Ranked Search, Similarity Matching.*

1. INTRODUCTION

Cloud computing is a term used to describe a set of IT services that are provided to a customer over a network on a leased basis and with the ability to scale up or down their service requirements. Clouds are large pools of easily usable and accessible virtualized resources. These resources can be dynamically reconfigured to adjust to a variable load (scale), permitting optimum resource utilization. It is a pay-per-use model in which the Infrastructure Provider

by means of customized Service Level Agreements (SLAs)[1] offers guarantees typically exploiting a pool of resources. Organizations and individuals can benefit from mass computing and storage centers, provided by large companies with stable and strong cloud schemas.Security concerns is the major challenges in cloud computing. The hardware and software security mechanisms like firewalls etc. have been used by cloud provider. These solutions are not sufficient to protect data in cloud from unauthorized users because of low degree of transparency [2]. Since the cloud user and the cloud provider are in the different trusted domain, the outsourced data may be exposed to the vulnerabilities [2] [3]. Thus, before storing the valuable data in cloud, the data needs to be encrypted [5]. Data encryption assures the data confidentiality and integrity. To preserve the data privacy we need to design a searchable algorithm that works on encrypted data [6]. To protect data privacy, confidentiality, and data security, sensitive data like personal health records, e-mails, tax documents, photo albums, financial transactions, and so on, have to be encrypted by data owners before outsourcing to the public cloud [7]. However, the traditional plaintext keyword search data utilization service is obsolete. Downloading all the information and decrypting at the information user side is trivially impractical duo to large amount of bandwidth cost is required in cloud scale systems. The data can be easily searched and utilized otherwise no purpose of storing information in the cloud. Thus, exploring effective and secure search over encrypted cloud information is of most importent. This is a very challenging issue; it degrades performance of system usability and size. It is very difficult to meet the requirements of system usability, performance and scalability by considering the huge number of on-demand information users and large number of outsourced information documents in the cloud. To meet effective information retrieval, the large amount of documents demands the cloud server to perform relevance score as a result, instead returning all result documents. Such ranking system facilitates information users to find the most relevant information rapidly, rather than burdensome sorting through every match in the information collection [8]. However, this will result in anhuge cost in terms of data, ease of use. For example, the existing models for keyword-based information retrieval, which are regularly used on the plaintext information, cannot be applied directly to the encrypted data. Download all data in the cloud and to decrypt locally is obviously impractical. To solve the above problem, researchers have some general purpose solutions with fully homomorphic encryption or blind RAMs [9] constructed. These methods are not practical due to their high computational cost for both the cloud Sever and users. Proposed scheme to achieve flexible search sub-linear search time and deal with the deletion and insertion of documents.

RELATED WORK

Many searching techniques over encrypted cloud data have proposed. S.Deshpande [11] suggested a technique searching over encrypted cloud data using fuzzy keywords. They used Edit distance to quantify keyword similarity and developed two techniques on constructing fuzzy keyword sets to achieve optimized storage and representation

overheads. Cong wang et al. [12] Has proposed a method ranked keyword search over encrypted cloud data using keyword frequency and order preserving encryption. It supports only single keywords at a time. Is the keyword frequency deciding document file score. Rank given to every file based on the relevance score of that file. Top ranked files have sent to users instead all files. To enrich search functionality N. Cao et al. [13] Have proposed a scheme supporting conjunctive keywords search. It is privacy – preserving multi-keyword ranked search technique using symmetric encryption. M. Chou et al. [14] proposed a solution for fuzzy multi-keyword search over encrypted cloud data using privacy aware Bed Tree.They used a co-occurrence probability approach to identify useful multi-keywords for publishing data, documents and relevant fuzzy keyword sets constructed using edit distance. They constructed index tree for all data, documents, where each leaf node having the hash value of a keyword, one or two data vectors that represents n- gram of that keyword and bloom filters for each edit distance value.Chi Chen, has proposed a hierarchical clustering method to more search support semantics and the demand for fast passphrase -Search meet in a big data environment [15] .The proposed hierarchical approach clusters the documents on the basis of minimum relevance threshold , and then partitions The resulting hierarchical cluster is reached , therestriction on the maximum size of the cluster [15] . In the search phase can achieve a linear computational complexity compared to an exponential increase in the size of document collection this approach. To verify the authenticity of the search results, a structure called minimum hash sub tree is designed in this paper. The proposed method has an advantage over the traditional method in the Rank Privacy as relevant documents.



Fig No 1. Hash Sub Tree Designed

2. SYSTEM MODEL

We considered a cloud computing system model having three different entities. Those are Data Owner, Cloud Service Provider and Data. The responsibility of each entity is as follows: Data Owner (DO): DO has a collection data documents DC= {d1, d2…, dm} with sensitive information to be outsourced to the cloud server. To provide data privacy, the documents are encrypted before outsourcing. DO creates a dictionary based on keywords extracted from the all m documents based on Term Frequency Inverted Document Frequency (TFIDF) [16] which is described in section 4. The dictionary includes synonyms of each keyword from the thesaurus [17]. The dictionaryishavingand keywords, and for each keyword may have t synonyms, so that the dictionary size is n × t. DO creates an index vector for each document based on the keywords extracted from the document. The size of the index vector is equal to the number of keywords in the dictionary that is

an. Each dimension in the index vector stores sum of the frequency of keyword and corresponding synonyms in the dictionary is denoted as term frequency (TF) in our system. Index vectors of all documents are encrypted before outsource to the cloud. DO create query vector based on keywords entered by Data user. To provide user privacy, query vector encrypted, as Trapdoor and send to Data user. The data owner sends search access control to the authorized data user.

Data users:

Data users are the users who accessing sensitive data from the cloud. The cloud server searches keywords or synonyms related to documents, which are interested to information user and sends to the information owner. The information user receives trapdoor and searches access control of data owner and sends trapdoor and access control to the cloud server to retrieve required information from the cloud.

A. Cloud Service Provider (CSP):

Cloud server receives encrypted data and encrypted index vectors from data owner and stores into data owner's cloud storage. Cloud server having the capability to take the information request from user and check the search access control of the user. It will retrieve the documents from cloud storage depending upon the privileges to access number of documents. To increase the data retrieval accuracy from cloud server, the top scored documents return to data user from the cloud server. The model for multi-keyword synonym query over encrypted cloud data.

B. Threat model:

The cloud server is measured as "honest-but-curious" [18] in our proposed approach. The cloud server follows the proposed method specification and also observes information in its cloud storage and data which are received from data user through the treating to learn extra information. We consider one threat model for our system with different attack capabilities that is as follows: Known ciphertext model: In this model, the cloud server knows only encrypted data and encrypted index vectors, which are outsourced from data owner.

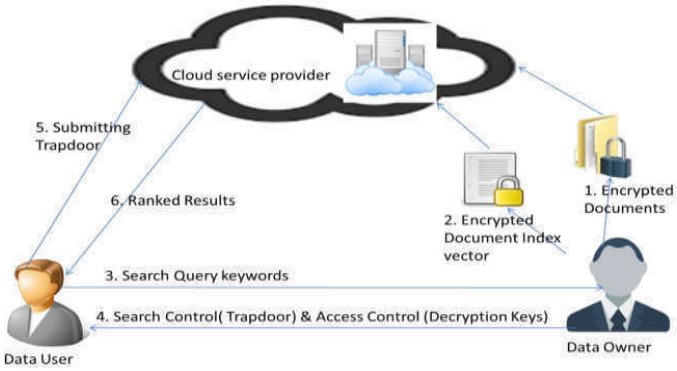


Fig No 2 System Model

3. PROPOSED SYSTEM

The first symmetrical searchable encryption (SSE) scheme and the search of the scheme is linear in the size of the data collection. Proposed formal security definitions for SSE and developed a system based on Bloom filter. It is proposed that two systems (SSE - 1 and 2) that the optimal search time is reached. Your SSE 1 scheme is secure against attacks Chosen- Keyword (CKA1) and SSE -2 is secure against adaptive chosen- keyword attacks (CKA2). These early works are single keyword Boolean search schemes that are very simple in terms of functionality. After plenty of plants have been proposed under different threat models to search various search functions, such as

ISBN 978-81-939386-1-4

single keyword search, similarity search more keyword Boolean search space and multi keyword search on place, etc. Multi - keyword Boolean search allows achieve the user to enter multiple query keywords to request appropriate documents. Among these works, combining keyword search systems give only the documents that contain all of the query keywords. Disjunctive Keyword Schemes return all documents that contain keywords proposed [19] .Predicate search schemes a subset of the query, both connecting divisive to support search. All these schemes More Keyword retrieve search results based on the presence of keywords, which can provide not acceptable result ranking functionality [20]. Proposed guide can achieve sublinear search time flexible and deal with the deleting and inserting documents. The safe kNN algorithm used to encrypt the index and query vectors, in the meantime accurate relevancy score calculation between encrypted index and query vectors [21] .Ensure to withstand various attacks in different threat models, build two secure search systems: the dynamic top k multi-keyword search scheme selected in the known ciphertext model, and improved dynamic top k multi- keyword space search procedure in the known background model. For our system, we choose the B-tree as indexing data structure to identify the match between search query and information documents. Specially, we use inner documents correspondence, i.e., the number of query keywords appearing in document, to finding the similarity of that document to the search query. Each document is transformed to a balanced B-tree according to the keywords and encrypted using CRSA. Whenever user wants to search, He creates a trapdoor for the keywords. Our aim is to build and analyze the performance of multiple keywords ranked search pattern using Commutative RSA algorithm and B-tree data structure for searchable index tree.

1. **Commutative Encryption (CRSA):**

The RSA cryptosystem is one of the best public key cryptography approaches. However, its overall robustness gets limited due to one way encryption and majority of existing RSA model suffer from reorder issues. Therefore, in order to make this system least complicated and more efficient, an approach called Commutative RSA has been proposed. In this scheme, the order in which encryption has been done would not affect the decryption if it is done in the same order. Encryption is the fixed method for making a communication private. With the many cryptographic approaches, our system follows the commutative RSA algorithm. The mathematical scheme for performing this encryption is described by a pseudo algorithm.

2. **BMS Tree Index Construction**:

In the process index tree construction, we generate node for each document in the document collection. These nodes are act as leaf nodes in the tree. The internal nodes are formed based on these leaf nodes. The index tree construction process is described in the algorithm 1. An example of BMS index tree for our scheme which is constructed on plaintext. The data structure of the node is defined as (ID, F, child [], DID), where ID is a unique id generated using GenID() function, F is index vector, child[] is pointers to children of the node and DID is a document ID. In the algorithm, we used two variables Current Node Collection and

ISBN 978-81-939386-1-4

Temp Node Collection to store collection of nodes. Current Node Collection stores the set of currently processing nodes which have no parents and Temp Node Collection stores set of newly formed nodes. Fu[i] always stores the biggest TF value of wiamong its children. The possible largest relevance score of its children is estimated using this technique.

Algorithm 1 Build BMS Index Tree(DC)

For each data document Ddid in DC do
Construct leaf node l for Ddid l.ID=GenID(), l.child[i]=null for i=1,…, b; l.DID=DID, and F[i]=TFDdid,ki for i=1,…, n;
Insert l to CurrentNodeCollection;
End for While the number of nodes in CurrentNodeCollection is more than 1 do
For each five of nodes u1, u2, u3, u4, and u5 in
CurrentNodeCollection do
Generate a parent node u for u1, u2, u3, u4, and u5with u.ID=GenID(), u.child[i] = uifor i = 1to 5; u.DID = 0, and D[i] = max{ui.F[j] for i=1 to 5} for each j=1 to n;
Insert u to TempNodeCollection;
End for
The remaining nodes (less than 5 nodes) in CurrentNodeCollection generate a parent node u like above;
Insert u to TempNodeCollection;
Replace CurrentNodeCollection with TempNodeCollection and then free the TempNodeCollection;
End while
Return only one node, left in the CurrentNodeCollection called the root node;

3. Search Process using DFST:

The search process of MSRQE scheme is the recursive function upon the BMS tree name as Depth First Search Technique algorithm. We create a result documents as RankedList, whose element is denoted as (Score, DID). Here, the score is the relevance score between Fdid and query vector Q, which is calculated using formula(1). The RankedList stores top k scored documents to query. The elements of RankedList are in descending order according to score function during the search process. The DFST algorithm is presented in algorithm 2. Kth score is a smallest relevance score in RankedList.

Algorithm 2 DFST(Index Tree Node u)

If the node u is not a leaf node then
If Score(Fu, Q) >kth score then
Sort the children of u in descending order according to scores of children
For i=1 to the number of children of u do
GDFS(u.child[i]);
End for
Else
Return;
End if
Else
If Score(Fu, Q) >kth score then
Delete the element with a smallest relevance score from
RankedList;
Insert a new element (Score (Fu, Q), u.ID) and sort all elements of RankedList in descending order;
End if
Return;
End if.

Designing searchable encryption schemes, the structures obtained with sub- linear search time, In particular, dynamic search is simple, highly parallel and can easily handle updates. Specifically, for n documents on m keywords and with p cores (processors) indicated available and it supports complex

queries and multi- client settings using SSE as a black box. To ensure that our data structures and related business activities which is support for dynamic databases and it support terabyte -Scale databases in these richer / complex encrypted search settings. The dynamic expansion maintains the optimum index size and only basic size information.

4. MRSE FRAMEWORK

For easy presentation, operations on the data documents are not shown in the framework since the data owner could easily employ the traditional symmetric key cryptography to encrypt and then outsource data. With focus on the index and query, the MRSE system consists of four algorithms as follows

1. **Setup(ℓ)** Taking a security parameter ℓ as input, the data owner outputs a symmetric key as SK.

2. **BuildIndex(F, SK**) Based on the dataset F, the data owner builds a searchable index I which is encrypted by the symmetric key SK and then outsourced to the cloud server. After the index construction, the document collection can be independently encrypted and outsourced.

3. **Trapdoor(fW**) With t keywords of interest in fW as input, this algorithm generates a corresponding trapdoor TfW.

4. **Query(TfW, k, I**) When the cloud server receives a query request as (TfW, k), it performs the ranked search on the index I with the help of trapdoor TfW, and finally returns FfW, the ranked id list of top-k documents sorted by their similarity with fW.

The representative privacy guarantee in the related literature, such as searchable encryption, is that the server should learn nothing but search results. With this general privacy description, we explore and establish a set of strict privacy requirements specifically for the MRSE framework. As for the data privacy, the data owner can resort to the traditional symmetric key cryptography to encrypt the data before outsourcing, and successfully prevent the cloud server from prying into the outsourced data.

5. RESULTS AND DISCUSSION

The proposed scheme, data users can achieve different requirements on search precision of privacy by the standard deviation of adjustment that can be treated as a compensation parameter. The comparison of systems with a recent work that achieves high search efficiency. BDMRS scheme calls the search results by exact calculation of document vector and query vector. Thus, top- k search accuracy of BDMRS scheme is 100 %. But based and similarity Multi- keyword square search pattern, the basic scheme in suffering from loss of precision due to the accumulation of sub-vectors with the index construction . The test is repeated 16 times, and the average accuracy of 91 %. During the search, when the relevance of the node is greater than the minimum relevance in results Rlist, examines the cloud server, the children of the node; otherwise it returns. So many nodes not accessed during a real search. We denote the number of leaf nodes that contain one or more keywords in the query. It is generally greater than the number of documents required k, but far less than the cardinality of the document collection n. As a balanced binary tree, the height of the index n is log will be

maintained, and the complexity of the calculation is ranked relevance O (m).

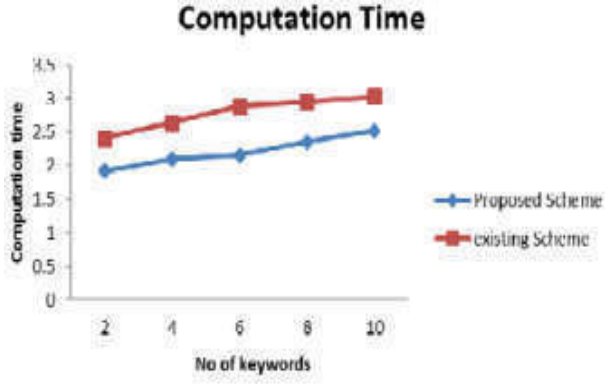


Figure 3: Time Comparison

The graph the comparison of the search computation time in seconds of our proposed system against the RSA based system. For two keywords search, the time taken by the RSA based scheme is approximately 2.5 seconds, whereas our proposed system takes approximately 0.5 seconds less. As the number of keywords increased for search, the computation time for search also increases linearly in both schemes. But CRSA based scheme is found to perform better. Thus it is evident that encryption algorithm CRSA with B Tree as index tree performs better than RSA and B tree Combination

6. CONCLUSION

We intend to provide feasible solutions for multi-keyword synonym ranked query problems over encrypted cloud data while preserving strict system wise privacy in cloud computing paradigm. The first one multi-keyword search, the second synonym based search, third similarity ranked search and the last is efficient data retrieval with BMS tree and DFST searching algorithm. Our example illustration further shows efficient and accurate top k documents retrieval of proposed scheme with sub-linear time complexity. Multi rank keyword search scheme is proposed, which not only supports true multi-keyword search on space, but also the dynamic deletion and insertion of documents. We build a special keyword balanced binary tree as the index. In addition, the search process may be performed in parallel to reduce the time, cost. The security of the system is protected against two threat models through secure top-k retrieval algorithm. The experimental results show the effectiveness of our proposed scheme. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given, and experiments on the real-world dataset shows our proposed scheme introduces low overhead on both computation and communication.

7. FUTURE WORK

The future work would concentrate on using Elliptic Curve Cryptography (ECC) encryption technique for better performance. Further, we intend to analyze the behavior of our proposed system(s) for multiuser environment.The dynamic operation such as updating and deletion has to assume with privacy and security policies.

8. REFERENCES

[1] K. Ren, C. Wang, Q. Wang et al., "Security challenges for the public cloud," IEEE Internet Computing, vol. 16, no. 1, pp. 69–73, 2012.

[2] Cloud Security Alliance, 'Security Guidance for Critical Areas of Focus in Cloud Computing,' http://www.cloudsecurityalliance.org, 2009.

[3] R. Brinkman, 'Searching in encrypted data,' in University of Twente, PhD thesis, 2007.

[4] S. Kamara and K. Lauter, 'Cryptographic cloud storage,' in RLCPS, January 2010, LNCS. Springer, Heidelberg.

[5] A. Singhal, 'Modern information retrieval: A brief overview,' IEEE Data Engineering Bulletin, vol. 24, no. 4, pp. 35–43, 2001.

[6] W. K. Wong, D. W. Cheung, B. Kao, and N. Mamoulis, 'Secure knn computation on encrypted databases,' in Proc. of SIGMOD, 2009.

[7] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," Proc. 14th Int'l Conf. Financial Cryptography and Data Security, Jan. 2010.

[8] Zhihua Xia, Xinhui Wang, Xingming Sun, and Qian Wang, "A Secure and Dynamic Multi-keyword RankedSearch Scheme over Encrypted Cloud Data," IEEE Transactions on Parallel and Distributed Systems, 2015

[9] KawserWazedNafi, TonnyShekhaKar, SayedAnisulHoque, Dr. M. M. A Hashem, "A Newer User Authentication, File encryption and Distributed Server Based Cloud Computing security architecture "Lecturer, Stamford University, Bangladesh, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 10, 2012

[10] Chinua Xia, Xinhui Wang," A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data", DOI 10.1109/TPDS. 2401003, IEEE Transactions on Parallel and Distributed Systems, 2015.

[11] S. Deshpande, "Fuzzy keyword search over encrypted data in cloud computing",World Journal of Science and Technology, vol. 2, no. 10, (2013).

[12] D.X.Song,D. Wagner and A.Perrig,"Practical techniques for searches on encrypted data, In Security and Privacy", 2000. S&P 2000,IEEE, (2000).

[13] N. Cao,C. Wang, M. Li, K. Ren, and W. Lou,"Privacy-preserving multi-keyword ranked search over encrypted cloud data", INFOCOM, 2011 Proceedings IEEE, (2011).

[14] C. Wang, KuiRen, Shucheng Yu, Urs, K.M.R "Achieving usable and privacy-assured similarity search over outsourced cloud data", INFOCOM, 2012 Proceedings IEEE, (2012)

[15] Chi Chen, Xiaojie Zhu, "An Efficient Privacy-Preserving Ranked Keyword Search Method", Member, IEEE, IEEE DOI 10.1109/TPDS.2425407, IEEE Transactions on Parallel and Distributed Systems, 2015.

[16] Yi Yang, Hongwei Li, Wenchao Liu, Haomiao Yao, Mi Wen," Secure Dynamic Searchable Symmetric Encryption with Constant Document Update Cost", School of Computer Science and Engineering, University of Electronic Science and Technology of China, Globecom - Communication and Information System Security Symposium, 2014.

[17] Chi Chen, Xiaojie Zhu, "An Efficient Privacy-Preserving Ranked Keyword Search Method", Member, IEEE, IEEE DOI 10.1109/TPDS.2425407, IEEE Transactions on Parallel and Distributed Systems, 2015.

[18] Hongwei Li, Dongxiao Liu, Kun Jia, and XiaodongLinss"Achieving Authorized and Ranked MultikeywordSearch over Encrypted Cloud Data" School of Computer Science and Engineering, University of Electronic Science and Technology of

China. IEEE ICC - Communication and Information Systems Security Symposium,2015

[19] Zhangjie Fu, KuiRen, JiangangShu, Xingming Sun "Enabling Personalized Search over Encrypted OutsourcedData with Efficiency Improvement", DOI 10.1109/TPDS.2506573, IEEE Transactions on Parallel and Distributed System, 2015

[20] Wenhai Sun, Bing Wang, Ming Cao,"Privacy-preserving Multi-keyword Text Search in the Cloud SupportingSimilarity-based Ranking "asia ccs'13, May 8– 10, Hangzhou, China. Copyright 2013 acm 978-1-4503- 1767-2/13/05, 2013.

[21] KawserWazedNafi, TonnyShekhaKar, SayedAnisulHoque, Dr. M. M. A Hashem, "A Newer UserAuthentication, File encryption and Distributed Server Based Cloud Computing security architecture "Lecturer,Stamford University, Bangladesh, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 10, 2012 [12] Chinua Xia, Xinhui Wang," A Secure.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Oruta: Privacy-Maintaining Public Scrutinize for Shared Data in the Cloud

Thana Mallika
Asst.prof
Dept. CSE
MRCE, Hyderabad.
Thanna.mallika@gmail.com

Vaddi Devipriya
Asst.prof
Dept. CSE
MRCE, Hyderabad.
vaddi.devipriya@gmail.com

Rudra koteswaramma
Asst.prof
Dept. CSE
MRCE , Hyderabad.
rudraeswari@gmail.com

Abstract

cloud information services, it's commonplace for information to be not solely hold on within thecloud, however conjointly shared across multiple users. sadly, the integrity of cloud information is subject to skepticism because of the existence of hardware/software failures and human errors. many mechanisms aredesigned to permit each information homeowners and public verifiers to expeditiously audit cloud informationintegrity while not retrieving the complete information from the cloud server. However, public auditing on the integrity of shared information with these existing mechanisms can inevitably reveal confidential information—identity privacy—to public verifiers. during this paper, we tend to propose a completely unique privacy-preserving mechanism that supports public auditing on shared information hold on within the cloud. above all, we tend toexploit ring signatures to work out verification information required to audit the correctness of shared information. With our mechanism, the identity of the signer on every block in shared information is unbroken personal from public verifiers, United Nations agency area unit able to expeditiously verify shared information integrity while notretrieving the complete file. additionally, our mechanism is ready to perform multiple auditing tasks at the same time rather than collateral them one by one. Our experimental results demonstrate the effectiveness and potencyof our mechanism once auditing shared information integrity

Index Terms—Public scrutinizing, privacy-maintaining, shared data, cloud computing

Ç

1 INTRODUCTION

CLOUD service suppliers supply users economical and scal- ready information storage services with a far lower marginal cost than ancient approaches [2]. it's routine for users to leverage cloud storage services to share information with others in a very cluster, as information sharing becomes a regular feature in most cloud storage offerings, together with Dropbox, iCloud and Google Drive.

The integrity of information in cloud storage, however, is subject to skepticism and scrutiny, as information keepwithin the cloud will simply be lost or corrupted thanks to the inevitable hardware/ computer code failures and human errors [3], [4]. to create this matter even worse, cloud service suppliers is also reluc- tant to tell users regarding these information errors so as to keep up the name of their services and avoid losing profits [5]. Therefore, the integrity of cloud information ought to be verified before any information utilization, like search or com- putation over cloud information [6].

The traditional approach for checking information correctness is to retrieve the complete information from the cloud, so ver- ify information integrity by checking the correctness of signatures (e.g., RSA [7]) or hash values (e.g., MD5 [8]) of the complete information. Certainly, this typical approach is ready to suc- cessfully check the correctness of cloud information. However

the potency of victimization this ancient approach on cloud knowledge is doubtful[9] The main reason is that the scale of cloud knowledge is massive normally.Downloading the whole cloud knowledge to verify knowledge integrity can value or maybe waste users amounts of computation and communication resources, particularly once knowledge are corrupted within the cloud. Besides, several uses of cloud knowledge (e.g., data processing and machine learning) don't essentially wantusers to transfer the entire cloud knowledge to native devices [2]. it's as a result ofcloud suppliers, like Amazon, offers users computation services directly on large-scale knowledge that already existed withinthe cloud.

Recently, several mechanisms [9], [10], [11], [12], [13], [14], [15], [16], [17] are projected to permit not solely {a knowledge|a knowledge|an information} owner itself however conjointly a public admirer to with efficiencyperform integrity checking while not downloading the entire data from the cloud, that is cited as public auditing [5]. In these mechanisms, knowledge is split into several little blocks, wherever every block is severally signed by the owner; and a random combination of all the blocks rather than the entire knowledge is retrieved throughout integrity checking [9]. A public admirer might be a knowledge user (e.g., researcher) United Nations agency would really like to utilize the own- er's knowledge via the cloud or a third-party auditor (TPA) United Nations agency willgive professional integrity checking services [18]. Mov- ing a revolution, Wang et al. designed a complicated auditing mechanism [5] (named as WWRL during thispaper), so throughout public auditing on cloud knowledge, the content of personalknowledge happiness to a private user isn't dis- closed to any public verifiers. sadly, current pub- lic auditing solutions mentioned higher than solely target personal knowledge within the cloud [1].

We believe that sharing knowledge among multiple users is maybe one in every ofthe foremost participating options that motivates cloud storage. Therefore, it'sconjointly necessary to make sure the integrity of shared knowledge within thecloud is correct. Existing

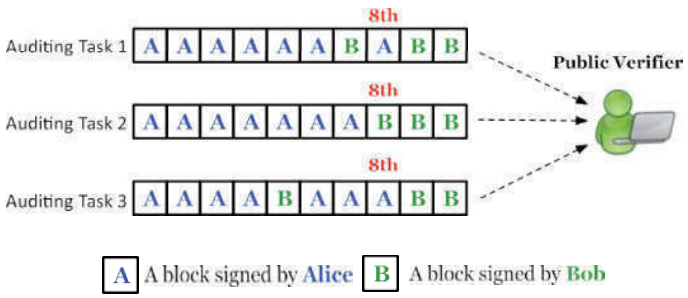


Fig. 1. Alice and Bob share a data file in the cloud, and a public verifier audits shared data integrity with existing mechanisms.

public auditing mechanisms can actually be extended to verify shared data integrity [1], [5], [19], [20]. However, a new significant privacy issue introduced in the case of shared data with the use of existing mechanisms is the leakage of *identity privacy* to public verifiers [1].

For instance, Alice and Bob work together as a group and share a file in the cloud (as presented in Fig. 1). The shared file is divided into a number of small blocks, where each block is independently signed by one of the two users with existing public auditing solutions (e.g., [5]). Once a block in this shared file is modified by a user, this user needs to sign the new block using his/her private key. Eventually, different blocks are signed by different users due to the modification introduced by these two different users. Then, in order to correctly audit the integrity of the entire data, a public verifier needs to choose the appropriate public key for each block (e.g., a block signed by Alice can only be correctly verified by Alice's public key). As a result, this public verifier will inevitably learn the identity of the signer on each block due to the unique binding between an identity and a public key via digital certificates under public key infrastructure (PKI).

Failing to preserve identity privacy on shared data during public auditing will reveal significant confidential information (e.g., which particular user in the group or special block in shared data is a more valuable target) to public verifiers. Specifically, as shown in Fig. 1, after performing several auditing tasks, this public verifier can first learn that Alice may be a more important role in the group because most of the blocks in the shared file are always signed by Alice; on the other hand, this public verifier can also easily deduce that the eighth block may contain data of a higher value (e.g., a final bid in an auction), because this block is frequently modified by the two different users. In order to protect these confidential information, it is essential and critical to preserve identity privacy from public verifiers during public auditing.

In this paper, to solve the above privacy issue on shared data, we propose Oruta,[1] a novel privacy-preserving public auditing mechanism. More specifically, we utilize ring signatures [21] to construct homomorphic authenticators [10] in Oruta, so that a public verifier is able to verify the integrity of shared data without retrieving the entire data—while the identity of the signer on each block in shared data is kept private from the public verifier. In addition, we further extend our mechanism to support batch auditing, which can perform multiple auditing tasks simultaneously and improve the efficiency of verification for multiple auditing tasks. Meanwhile, Oruta is compatible with random masking [5], which has been utilized in WWRL and can preserve data privacy from public verifiers. Moreover, we also leverage index hash tables from a previous public auditing solution [15] to support dynamic data. A high-level comparison among Oruta and existing mechanisms is presented in Table 1.

1. Oruta stands for "One Ring to Rule Them All."

TABLE 1
Comparison among Different Mechanisms

	PDP [9]	WWRL [5]	Oruta
Public Auditing	√	√	√
Data Privacy	×	√	√
Identity Privacy	×	×	√

The remainder of this paper is organized as follows. In Section 2, we present the system model, threat model and design objectives. In Section 3, we introduce cryptographic primitives used in Oruta. The detailed design and security analysis of Oruta are presented in Section 4 and Section 5. In Section 6, we evaluate the performance of Oruta. Finally, we briefly discuss related work in Section 7, and conclude this paper in Section 8.

2 Problem Statement

System Model

As illustrated in Fig. 2, the system model in this paper involves three parties: the cloud server, a group of users and a public verifier. There are two types of users in a group: the original user and a number of group users. The original user initially creates shared data in the cloud, and shares it with group users. Both the original user and group users are members of the group. Every member of the group is allowed to access and modify shared data. Shared data and its verification metadata (i.e., signatures) are both stored in the cloud server. A public verifier, such as a third-party auditor providing expert data auditing services or a data user outside the group intending to utilize shared data, is able to publicly verify the integrity of shared data stored in the cloud server.

When a public verifier wishes to check the integrity of shared data, it first sends an auditing challenge to the cloud server. After receiving the auditing challenge, the

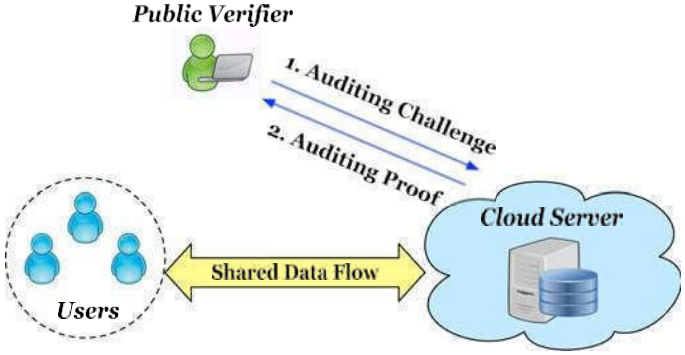


Fig. 2. Our system model includes the cloud server, a group of users and a public verifier.

cloud server responds to the public verifier with an auditing proof of the possession of shared data. Then, this public verifier checks the correctness of the entire data by verifying the correctness of the auditing proof. Essentially, the process of public auditing is a challenge-and-response protocol between a public verifier and the cloud server [9].

Threat Model

Integrity Threats. Two kinds of threats related to the integrity of shared data are possible. First, an adversary may try to corrupt the integrity of shared data. Second, the cloud service provider may inadvertently corrupt (or even remove) data in its storage due to hardware failures and human errors. Making matters worse, the cloud service provider is economically motivated, which means it may be reluctant to inform users about such corruption of data in order to save its reputation and avoid losing profits of its services.

Privacy Threats. The identity of the signer on each block in shared data is private and confidential to the group. During the process of auditing, a public verifier, who is only allowed to verify the correctness of shared data integrity, may try to reveal the identity of the signer on each block in shared data based on verification metadata. Once the public verifier reveals the identity of the signer on each block, it can easily distinguish a high-value target (a particular user in the group or a special block in shared data) from others.

Design Objectives

Our mechanism, Oruta, should be designed to achieve following properties: (1) *Public Auditing*: A public verifier is able to publicly verify the integrity of shared data without retrieving the entire data from the cloud. (2) *Correctness*: A public verifier is able to correctly verify shared data integrity. (3) *Unforgeability*: Only a user in the group can generate valid verification metadata (i.e., signatures) on shared data. (4) *Identity Privacy*: A public verifier cannot distinguish the identity of the signer on each block in shared data during the process of auditing.

Possible Alternative Approaches

To preserve the identity of the signer on each block during public auditing, one possible alternative approach is to ask all the users of the group to share a *global private key* [22], [23]. Then, every user is able to sign blocks with this global private key. However, once one user of the group is compromised or leaving the group, a new global private key must be generated and securely shared among the rest of the group, which clearly introduces huge overhead to users in terms of key management and key distribution. While in our solution, each user in the rest of the group can still utilize its own private key for computing verification metadata without generating or sharing any new secret keys.

Another possible approach to achieve identity privacy, is to add a trusted proxy between a group of users and the cloud in the system model. More concretely, each member's data is collected, signed, and uploaded to the cloud by this trusted proxy, then a public verifier can only verify and learn that it is the proxy signs the data, but cannot learn the identities of group members. Yet, the security of this method is threatened by the single point failure of the proxy. Besides, sometimes, not all the group members would like to trust the same proxy for generating signatures and uploading data on their behalf. Utilizing group signatures [24] is also an alternative option to preserve identity privacy. Unfortunately, as shown in our recent work [25], how to design an efficient public auditing mechanism based on group signatures remains open.[2]

Trusted Computing offers another possible alternative approach to achieve the design objectives of our mechanism. Specifically, by utilizing direct anonymous attestation [26], which is adopted by the Trusted Computing Group as the anonymous method for remote authentication in trusted platform module, users are able to preserve their identity privacy on shared data from a public verifier. The main problem with this approach is that it requires all the users using designed hardware, and needs the cloud provider to move all the existing cloud services to the trusted computing environment, which would be costly and impractical.

3 Preliminaries

In this section, we briefly introduce cryptographic primitives and their corresponding properties that we implement in Oruta.

Bilinear Maps

Let G_1, G_2 and G_T be three multiplicative cyclic groups of prime order p, g_1 be a generator of G_1, and g_2 be a generator of G_2. A bilinear map e is a map $e: G_1 \times G_2 \rightarrow G_T$ with the following properties:

- *Computability*: there exists an efficiently computable algorithm for computing map e.
- *Bilinearity*: for all $u \in G_1$, $v \in G_2$ and $a, b \in Z_p$, $e(u^a, v^b) = e(u, v)^{ab}$.
- *Non-degeneracy*: $e(g_1, g_2) \neq 1$.

Bilinear maps can be generally constructed from certain elliptic curves [27]. Readers do not need to learn the technical details about how to build bilinear maps from certain elliptic curves. Understanding the properties of bilinear maps described above is sufficient enough for readers to access the design of our mechanism.

Security Assumptions

The security of our proposed mechanism is based on the two following assumptions:

Computational Co-Diffie-Hellman (Co-CDH) Problem. Let $a \in Z_p^*$, given $g_2, g^a \in G_2$ and $h \in G_1$ as input, output $h^a \in G_1$.

Definition 1 (Computational Co-Diffie-Hellman Assumption). *The advantage of a probabilistic polynomial time*

2. The direct leverage of group signatures in an public auditing mechanism makes the size of verification metadata extremely huge, which is much larger than the size of data itself. See [25] for details.

algorithm A *in solving the Co-CDH problem on* $\delta G_1; G_2 \rho$ *is defined as*

$$\mathrm{AdvCoCDH}_A = \Pr[A(g_2; g_2^a; h) = h^a : a \in_R Z_p^*; h \in_R G_1];$$

where the probability is over the choice of a *and* h*, and the coin tosses of* A*. The Co-CDH assumption means, for any probabilistic polynomial time algorithm* A*, the advantage of it in solving the Co-CDH problem on* $\delta G_1; G_2 \rho$ *is negligible.*

For the ease of understanding, we can also say solving the Co-CDH problem on $\delta G_1; G_2 \rho$ is or computationally infeasible or hard under the Co-CDH assumption.

Discrete Logarithm (DL) Problem. Let $a \in Z_p^*$, given $g; g^a \in G_1$ as input, output a.

Definition 2 (Discrete Logarithm Assumption). *The advantage of a probabilistic polynomial time algorithm* A *in solving the DL problem in* G_1 *is defined as*

$$\mathrm{AdvDL}_A = \Pr[A(\delta g; g^a \rho) = a : a \in_R Z_p^*];$$

where the probability is over the choice of a*, and the coin tosses of* A*. The DL Assumption means, for any probabilistic polynomial time algorithm* A*, the advantage of it in solving the DL problem in* G_1 *is negligible.*

Ring Signatures

The concept of ring signatures was first proposed by Rivest et al. [28] in 2001. With ring signatures, a verifier is convinced that a signature is computed using one of group members' private keys, but the verifier is not able to determine which one. More concretely, given a ring signature and a group of d users, a verifier cannot distinguish the signer's identity with a probability more than $1=d$. This property can be used to preserve the identity of the signer from a verifier.

The ring signature scheme introduced by Boneh et al. [21] (referred to as BGLS in this paper) is constructed on bilinear maps. We will extend this ring signature scheme to construct our public auditing mechanism.

Homomorphic Authenticators

Homomorphic authenticators (also called homomorphic verifiable tags) are basic tools to construct public auditing mechanisms [1], [5], [9], [10], [12], [15]. Besides unforgeability (i.e., only a user with a private key can generate valid signatures), a homomorphic authenticable signature scheme, which denotes a homomorphic authenticator based on signatures, should also satisfy the following properties:

Let $\delta pk; sk\rho$ denote the signer's public/private key pair, s_1 denote a signature on block $m_1 \in Z_p$, s_2 denote a signature on block $m_2 \in Z_p$.

- *Blockless verifiability:* Given s_1 and s_2, two random values a_1, $a_2 \in Z_p$ and a block $m' = a_1 m_1 \rho a_2 m_2 \in Z_p$, a verifier is able to check the correctness of block m' without knowing block m_1 and m_2.
- *Non-malleability:* Given s_1 and s_2, two random values a_1, $a_2 \in Z_p$ and a block $m' = a_1 m_1 \rho a_2 m_2 \in Z_p$, a user, who does not have private key sk, is not able to generate a valid signature s' on block m' by linearly combining signature s_1 and s_2.

Blockless verifiability allows a verifier to audit the correctness of data stored in the cloud server with a special block, which is a linear combination of all the blocks in data. If the integrity of the combined block is correct, then the verifier believes that the integrity of the entire data is correct. In this way, the verifier does not need to download all the blocks to check the integrity of data. Non-malleability indicates that an adversary cannot gen- erate valid signatures on arbitrary blocks by linearly com- bining existing signatures.

4 New Ring Signature Scheme

Overview

As we introduced in previous sections, we intend to utilize ring signatures to hide the identity of the signer on each block, so that private and sensitive information of the group is not disclosed to public verifiers. However, traditional ring signatures [21], [28] cannot be directly used into public auditing mechanisms, because these ring signature schemes do not support blockless verifiability. Without blockless verifiability, a public verifier has to download the whole data file to verify the correctness of shared data, which consumes excessive bandwidth and takes very long verification times.

Therefore, we design a new homomorphic authenticable ring signature (HARS) scheme, which is extended from a classic ring signature scheme [21]. The ring signatures generated by HARS are not only able to preserve identity privacy but also able to support blockless verifiability. We will show how to build the privacy-preserving public auditing mechanism for shared data in the cloud based on this new ring signature scheme in the next section.

Construction of HARS

HARS contains three algorithms: KeyGen, RingSign and RingVerify. In KeyGen, each user in the group generates his/her public key and private key. In RingSign, a user in the group is able to generate a signature on a block and its block identifier with his/her private key and all the group members' public keys. A block identifier is a string that can distinguish the corresponding block from others. A verifier is able to check whether a given block is signed by a group member in RingVerify. Details of this scheme are described in Fig. 3.

Security Analysis of HARS

Now, we discuss security properties of HARS, including correctness, unforgeability, blockless verifiability, non-malleability and identity privacy.

Theorem 1. *Given any block* m*, its block identifier* id*, and its ring signature* $s = \delta s_1; \ldots; s_d \rho$*, a verifier is able to correctly check the integrity of this block under HARS.*

Proof. Based on properties of bilinear maps, the correctness of this scheme can be proved as follows:

Let $\mathbb{G}_1$, $\mathbb{G}_2$ and $\mathbb{G}_T$ be multiplicative cyclic groups of order p, g_1 and g_2 be generators of $\mathbb{G}_1$ and $\mathbb{G}_2$ respectively. Let $e : \mathbb{G}_1 \times \mathbb{G}_2 \to \mathbb{G}_T$ be a bilinear map, and $\psi : \mathbb{G}_2 \to \mathbb{G}_1$ be a computable isomorphism with $\psi(g_2) = g_1$. There is a public map-to-point hash function H_1: $\{0,1\}^* \to \mathbb{G}_1$, which can map a string $\{0,1\}^*$ into an element of $\mathbb{G}_1$ (i.e., an point on an elliptic curve). The total number of users in the group is d. The global parameters are $(e, \psi, p, \mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T, g_1, g_2, H_1, d)$.

KeyGen. For a user u_i, he/she randomly picks $x_i \xleftarrow{R} \mathbb{Z}_p$ and computes $w_i = g_2^{x_i} \in \mathbb{G}_2$. Then, user u_i's public key is $\mathbf{pk}_i = w_i$ and his/her private key is $\mathbf{sk}_i = x_i$.

RingSign. Given all the d users' public keys $(\mathbf{pk}_1, ..., \mathbf{pk}_d) = (w_1, ..., w_d)$, a block $m \in \mathbb{Z}_p$, the identifier of this block id and the private key $\mathbf{sk}_s$ for some s, user u_s randomly chooses $a_i \in \mathbb{Z}_p$ for all $i \neq s$, where $i \in [1, d]$, and let $\sigma_i = g_1^{a_i}$. Then, he/she computes

$$\beta = H_1(id) g_1^m \in \mathbb{G}_1, \tag{1}$$

and sets

$$\sigma_s = \left(\frac{\beta}{\psi(\prod_{i \neq s} w_i^{a_i})} \right)^{1/x_s} \in \mathbb{G}_1. \tag{2}$$

The ring signature of block m is $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_d) \in \mathbb{G}_1^d$.

RingVerify. Given all the d users' public keys $(\mathbf{pk}_1, ..., \mathbf{pk}_d) = (w_1, ..., w_d)$, a block m, an identifier id and a ring signature $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_d)$, a verifier first computes $\beta = H_1(id) g_1^m \in \mathbb{G}_1$, and then checks

$$e(\beta, g_2) \stackrel{?}{=} \prod_{i=1}^{d} e(\sigma_i, w_i). \tag{3}$$

If the above equation holds, then the given block m is signed by one of these d users in the group. Otherwise, it is not.

Fig. 3. Details of HARS.

$$\begin{aligned}
& e(\sigma_1, w_1) \cdot \prod_{i=2}^{d} e(\sigma_i, \overline{w}_i) \\
&= e\left(\frac{\beta}{\psi(\prod_{i \neq 1} w_i^{a_i})}, w_1 \right) \cdot \prod_{i=2}^{d} e(\sigma_i, \overline{w}_i) \\
&= e\left(\frac{\beta}{\psi(\prod_{i=2}^{d} w_i^{a_i})}, w_1 \right) \cdot \prod_{i=2}^{d} e(\sigma_i, \overline{w}_i) \\
&= e\left(\frac{\beta}{\psi(\prod_{i=2}^{d} g_2^{x_i a_i})}, g_2 \right) \cdot \prod_{i=2}^{d} e(g_1^{a_i}, g_2^{x_i}) \\
&= e\left(\frac{\beta}{\prod_{i=2}^{d} g_1^{x_i a_i}}, g_2 \right) \cdot \prod_{i=2}^{d} e(g_1^{a_i x_i}, g_2) \\
&= e\left(\frac{\beta}{\prod_{i=2}^{d} g_1^{a_i x_i}}, g_2 \right) \cdot \prod_{i=2}^{d} e(g_1^{a_i x_i}, g_2) \\
&= e(\beta, g_2),
\end{aligned}$$

where b is computed as $b = H_1(id) g_1^m$. □

Theorem 2. *For any adversary A, it is computationally infeasible to forge a ring signature under HARS, as long as the Co-CDH assumption on $(G_1; G_2)$ holds.*

Proof. We follow the security game defined in traditional ring signature schemes [21]. In the game, an adversary A is given all the d users' public key $(pk_1; \ldots; pk_d) = (w_1; \ldots; w_d)$, and is given access to the hash oracle and the ring signing oracle. The goal of adversary A is to output a valid ring signature on a pair of block/identifier $(m; id)$, where this pair of block/identifier $(m; id)$ has never been presented to the ring signing oracle. If adversary A achieves this goal, then it wins the game. And we can prove that if adversary A could win the game, then we can find an algorithm B to solve the Co-CDH problem on $(G_1; G_2)$ with a non-negligible advantage, which contradicts to the Co-CDH assumption we introduced in Section 3. So, we first assume adversary A is able to generate a forgery by the following security game simulated by algorithm B.

Initially, given $g_1^{ab} \in G_1$, $g_2^a \in G_2$, algorithm B randomly picks $x_2; \ldots; x_n$ from Z_p and sets $x_1 = 1$. Then, it sets $pk_i = w_i = (g_2^a)^{x_i}$. Adversary A is given the public keys $(pk_1; \ldots; pk_d) = (w_1; \ldots; w_d)$. Without loss of generality, we assume A can submit distinct queries, which means for every ring signing query on a block m and its identifier id, A has previously issued a hash query on block m and identifier id.

Hash Query. On a hash query issued by A, B flips a coin that shows 0 with probability p_c, and shows 1 otherwise. Then B randomly picks $r \in Z_p$. If the coin shows 0, B returns g_1^r to A, otherwise it returns $c g_2^r$. Since r is randomly selected from Z_p, and g_1 and $c g_2^a$ are both elements of cyclic group G_1, therefore, the distribution of $H(id)$ is identical to the distribution of $c g_2^{ar}$, which means A cannot distinguish the result of flipped coin from the result of the hash query.

Ring Signing Query. Suppose A issues a ring signing query on a block m and its identifier id. By the assumption of the game, a hash query has been issued by B on this pair of block/identifier $(m; id)$. If the coin B flipped for this hash query showed 0, then B fails and exits. Otherwise B has returned $H(id) g_1^m = c(g_2^a)^r$ for some r, which was randomly selected in the corresponding hash query. In this case, B chooses random $a_2; \ldots; a_d \in Z_p$, computes $a_1 = r - (a_2 x_2 + \cdots + a_d x_d)$, and returns the signature $\mathbf{s} = (g_2^{a_1}; g_2^{a_2}; \ldots; g_2^{a_d})$.

Eventually A outputs a forgery $\mathbf{s} = (s_1; \ldots; s_d)$ on block m and identifier id. Again by the assumption, a hash query has been issued on this pair of block/identifier $(m; id)$. If the coin flipped by B for this hash query did not show 0 then B fails. Otherwise, $H(id) g_1^m = g_1^{abr}$ for some r, which was randomly chosen by B in corresponding hash query, and B can output g_1^b by computing $\left(\prod_{i=1}^{d} s_i \right)^{1/r}$.

Clearly, given $g_1^{ab} \in G_1$, $g_2^a \in G_2$, algorithm B is able to output $g_1^b \in G_1$ via the above security game with an advantage of $sp_c^{q_s}(1 - p_c)$, if A is able to generate a forgery with an advantage of s. This advantage of algorithm B is maximized as $s/e \cdot (1 + q_s)$ when $p_c = q_s/(q_s + 1)$, where $e = \lim_{q_s \to \infty} (1 + 1/q_s)^{q_s}$.

According to [21], the Co-CDH problem can be solved by running two random instances of algorithm B. Therefore, if adversary A is able to generate a forgery of a ring signature with an advantage of s, then running two instances of algorithm B is able to solve Co-CDH

problem with an advantage of $\varepsilon(\lambda) \geq \varepsilon(\lambda)^2$. Clearly, if the advantage of adversary $\mathcal{A}$ of generating a forgery is non-negligible, then the advantage of solving the Co-CDH problem by running two instances of algorithm $\mathcal{B}$ is also non-negligible, which contradicts to the Co-CDH assumption in (G_1, G). Therefore, it is computationally infeasible for adversary $\mathcal{A}$ to generate a forgery. □

Theorem 3. *HARS is a homomorphic authenticable ring signature scheme.*

Proof. To prove HARS is a homomorphic authenticable ring signature scheme, we first prove that HARS is able to support blockless verifiability, which we defined in Section 3. Then we show HARS is also non-malleable.

Blockless Verifiability. Given all the d users' public keys $(pk_1, \ldots, pk_d) = (w_1, \ldots, w_d)$, two identifiers id_1 and id_2, two ring signatures $\mathbf{s}_1 = (s_{1,1}, \ldots, s_{1,d})$ and $\mathbf{s}_2 = (s_{2,1}, \ldots, s_{2,d})$, and two random values $y_1, y_2 \in Z_p$, a verifier is able to check the correctness of a combined block $m' = y_1 m_1 + y_2 m_2 \in Z_p$ without knowing block m_1 and m_2 by verifying:

$$e\left(H_1(id_1)^{y_1} H_1(id_2)^{y_2} g_1^{m'}, g_2\right) \stackrel{?}{=} \prod_{i=1}^{d} e\left(s_{1,i}^{y_1} \cdot s_{2,i}^{y_2}, w_i\right)$$

Based on Theorem 1, the correctness of the above equation can be proved as

$$\begin{aligned}
&e\left(H_1(id_1)^{y_1} H_1(id_2)^{y_2} g_1^{m'}, g_2\right) \\
&= e\left(H_1(id_1)^{y_1} g_1^{y_1 m_1}, g_2\right) \cdot e\left(H_1(id_2)^{y_2} g_1^{y_2 m_2}, g_2\right) \\
&= e\left(b_1, g_2\right)^{y_1} \cdot e\left(b_2, g_2\right)^{y_2} \\
&= \prod_{i=1}^{d} e\left(s_{1,i}, w_i\right)^{y_1} \cdot \prod_{i=1}^{d} e\left(s_{2,i}, w_i\right)^{y_2} \\
&= \prod_{i=1}^{d} e\left(s_{1,i}^{y_1} \cdot s_{2,i}^{y_2}, w_i\right).
\end{aligned}$$

If the combined block m' is correct, the verifier also believes that block m_1 and m_2 are both correct. Therefore, HARS is able to support blockless verifiability.

Non-Malleability. Meanwhile, an adversary, who does not have any user's private key, cannot generate a valid ring signature $\mathbf{s}'$ on the combined block $m' = y_1 m_1 + y_2 m_2$ by combining $\mathbf{s}_1$ and $\mathbf{s}_2$ with y_1 and y_2. Because if an element s'_i in $\mathbf{s}'$ is computed as $s'_i = s_{1,i}^{y_1} \cdot s_{2,i}^{y_2}$, the whole ring signature $\mathbf{s}' = (s'_1, \ldots, s'_d)$ cannot pass Equation (3) in *RingVerify*. The hardness of this problem lies in the fact that H_1 must be a one-way hash function (given every input, it is easy to compute; however, given the image of a random input, it is hard to invert).

Specifically, if block m_1 and m_2 are signed by the same user, for example, user u_s, then s_s can be computed as

$$s_s = s_{1,s}^{y_1} \cdot s_{2,s}^{y_2} = \left(\frac{b_1^{y_1} b_2^{y_2}}{\prod_{i \neq s}^{d} w_{1,i}^{y_1 a_{1,i}} \cdot w_{2,i}^{y_2 a_{2,i}}}\right)^{1/x_s}.$$

For all $i = 1, \ldots, s, \ldots, d$, $s'_i = s_{1,i}^{y_1} \cdot s_{2,i}^{y_2} = g_1^{(y_1 a_{1,i} + y_2 a_{2,i})}$, where $a_{1,i}$ and $a_{2,i}$ are random values. When this invalid ring signature $\mathbf{s}' = (s'_1, \ldots, s'_d)$ and the combined block m' are verified together with Equation (3), we have

$$\prod_{i=1}^{d} e\left(s'_i, w_i\right) = e\left(b_1^{y_1} b_2^{y_2}, g_2\right) = e(b', g_2),$$

which means it fails to pass the verification. The reason is that, although $b' = H(id_1)^{y_1} H(id_2)^{y_2} g_1^{m'}$, we can have $H(id_1)^{y_1} H(id_2)^{y_2} = H(id'_1)^{y_1} H(id'_2)^{y_2}$. Then, given a random value of $h = H(id_1)^{y_1} H(id_2)^{y_2}$ (due to y_1, y_2 are randoms), we can easily find an input id' so that $H(id') = h$, which contradicts to the assumption that H_1 is a one-way hash function.

If block m_1 and m_2 are signed by different users, for example, user u_s and user u_t, then s'_s and s'_t can be presented as

$$s'_s = \left(\frac{b_1^{y_1}}{\prod_{i \neq s} w_i^{y_1 a_{1,i}}}\right)^{1/x_s} \cdot g_1^{y_2 a_{2,s}},$$

$$s'_t = g_1^{y_1 a_{1,t}} \cdot \left(\frac{b_2^{y_2}}{\prod_{i \neq t} w_i^{y_2 a_{2,i}}}\right)^{1/x_t}.$$

For all $i = 1, \ldots, s, \ldots, t, \ldots, d$, $s'_i = s_{1,i}^{y_1} \cdot s_{2,i}^{y_2} = g_1^{(y_1 a_{1,i} + y_2 a_{2,i})}$, where $a_{1,i}$ and $a_{2,i}$ are random values. When this invalid ring signature $\mathbf{s}' = (s'_1, \ldots, s'_d)$ and the combined block m' are verified together with Equation (3), we have

$$\prod_{i=1}^{d} e\left(s'_i, w_i\right) = e\left(b_1^{1} b_2^{1}, g_2\right) = e(b', g_2),$$

which means it fails to pass the verification, due to the same reason explained in the previous case that H_1 is a one-way hash function. Therefore, an adversary cannot output valid ring signatures on combined blocks by combining existing signatures, which indicates that HARS is non-malleable. Since HARS is blockless verifiable and non-malleable, it is a homomorphic authenticable signature scheme. □

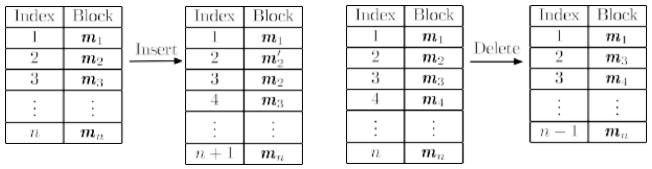
Theorem 4. *For any algorithm $\mathcal{A}$, any group U with d users, and a random user $u_s \in U$, the probability $Pr[\mathcal{A}(\mathbf{s}) = u_s]$ is at most $1/d$ under HARS, where $\mathbf{s}$ is a ring signature generated with user u_s's private key sk_s.*

Proof. For any $h \in G_1$ and any s, $1 \leq s \leq d$, the distribution $\{g_1^{a_1}, \ldots, g_1^{a_d} : a_i \in_R Z_p$ for $i = 1, \ldots, s, \ldots, d$, a_s chosen such that $\prod_{i=1}^{d} g_1^{a_i} = h\}$ is identical to the distribution $\{g_1^{a_1}, \ldots, g_1^{a_d} : \prod_{i=1}^{d} g_1^{a_i} = h\}$. Therefore, given $\mathbf{s} = (s_1, \ldots, s_d)$, the probability algorithm $\mathcal{A}$ distinguishes s_s, which indicates the identity of the signer, is at most $1/d$. Further explanations of this proof about identity privacy can be found in [21]. □

5 Public Auditing Mechanism

Overview

Using HARS and its properties we established in the previous section, we now construct Oruta, a privacy-preserving public auditing mechanism for shared data in the cloud. With Oruta, the public verifier can verify the integrity of shared data without retrieving the entire data. Meanwhile,



(a) After inserting block m'_2, all the identifiers after block m'_2 are changed

(b) After deleting block m_2, all the identifiers after block m_1 are changed

Fig. 4. Using indices as identifiers.

the identity of the signer on each block in shared data is kept private from the public verifier during the auditing.

Reduce Signature Storage

Another important issue we should consider in the construction of Oruta is the size of storage used for ring signatures. According to the generation of ring signatures in HARS, a block m is an element of Z_p and its ring signature contains d elements of G_1, where G_1 is a cyclic group with order p. It means a $|p|$-bit block requires a $d \times |p|$-bit ring signature, which forces users to spend a huge amount of space on storing ring signatures. It will be very frustrating for users, because cloud service providers, such as Amazon, will charge users based on the storage space they use.

To reduce the storage of ring signatures on shared data and still allow the public verifier to audit shared data efficiently, we exploit an aggregated approach from [10] to expand the size of each block in shared data into $k \times |p|$ bits. Specifically, a block in shared data is denoted as $m_j = (m_{j,1}, \ldots, m_{j,k}) \in Z_p^k$, for $1 \le j \le n$ and n is the total number of blocks. To generate a ring signature on block m_j with HARS, a user aggregates block $m_j = (m_{j,1}, \ldots, m_{j,k})$ as $\prod_{l=1}^{k} h_l^{m_{j,l}}$ instead of computing g^m in Equation (1), where $h_1, \ldots, h_k$ are random values of G_1. With the aggregation of a block, the length of a ring signature is only d/k of the length of a block. Similar methods to reduce the storage space of signatures can also be found in [15]. Generally, to obtain a smaller size of a ring signature than the size of a block, we choose $k > d$. As a trade-off, the communication cost of an auditing task will be increasing with an increase of k (analyzed in later section).

Support Dynamic Operations

To enable each user in the group to easily modify data in the cloud, Oruta should also support dynamic operations on shared data. A dynamic operation includes an insert, delete or update operation on a single block [9]. However, since the computation of a ring signature includes an identifier of a block (as presented in HARS), traditional methods, which only use the index of a block as its identifier (i.e., the index of block m_j is j), are not suitable for supporting dynamic operations on shared data efficiently.

The reason is that, when a user modifies a single block in shared data by performing an insert or delete operation, the indices of blocks that after the modified block are all changed (as shown in Fig. 4), and the changes of these indices require users, who are sharing the data, to re-compute the signatures of these blocks, even though the content of these blocks are not modified.

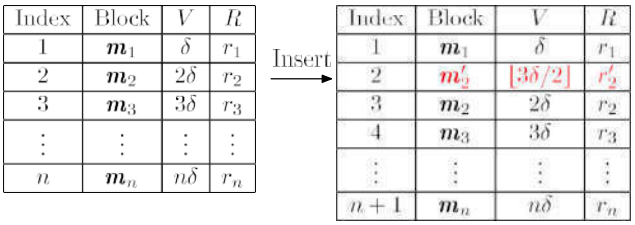


Fig. 5. Insert block m'_2 into shared data using an index hash table as identifiers.

By utilizing an index hash table [15], which is a data structure indexing each block based on its hash value, our mechanism can allow a user to efficiently perform a dynamic operation on a single block, and avoid this type of re-computation on other blocks. Different from [15], in our mechanism, an identifier from the index hash table is described as $id_j = \{v_j, r_j\}$, where v_j is denoted as the virtual index of block m_j, and r_j is a random generated by a collision-resistant hash function $H_2 : \{0,1\}^* \to Z_q$ with $r_j = H_2(m_j \| v_j)$. Here, q is a much smaller prime than p (e.g., $|q| = 80$ bits and $|p| = 160$ bits). Examples of different dynamic operations on shared data with our index hash tables are described in Figs. 5 and 6.

Specifically, the value of r generated by H_2 ensures that each block has a unique identifier (i.e., the probability that two blocks have the same value of r is negligible). The virtual indices are able to ensure that all the blocks in shared data are in the right order. For example, if $v_i < v_j$, then block m_i is ahead of block m_j in shared data. When shared data is created by the original user, the initial virtual index of block m_j is computed as $v_j = j \cdot d$, where d is a system parameter decided by the original user. If a new block m'_j is inserted, the virtual index of this new block m'_j is $v'_j = (v_{j-1} + v_j)/2$. Clearly, if block m_j and block m_{j+1} are both originally created by the original user, the maximum number of inserted blocks that is allowed between block m_j and block m_{j+1} is d. In this paper, we assume the value of d is always large enough to support a sufficient number of insertions between two blocks for the group, so that any two virtual indexes in the table will not be the same.

To support dynamic data without the above assumption, the combination of coding techniques and Oblivious RAM can be utilized as introduced in recent work [29]. Unfortunately, this proposed solution in [29] requires much more computation and communication overhead. Moreover, the property of public verifiability is inevitably sacrificed in [29].

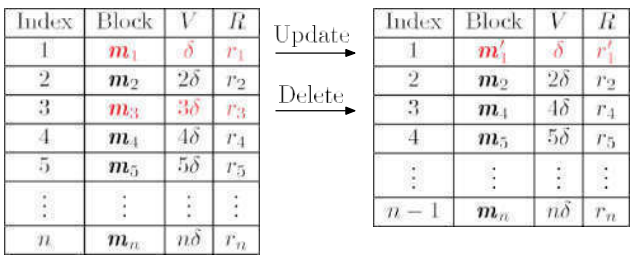


Fig. 6. Update block m_1 and delete block m_3 in shared data using an index hash table as identifiers.

Construction of Oruta

Now, we present the details of our public auditing mechanism. It includes five algorithms: KeyGen, SigGen, Modify, ProofGen and ProofVerify. In KeyGen, users generate their own public/private key pairs. In SigGen, a user (either the original user or a group user) is able to compute ring signatures on blocks in shared data by using its own private key and all the group members' public keys. Each user in the group is able to perform an insert, delete or update operation on a block, and compute the new ring signature on this new block in Modify. ProofGen is operated by a public verifier and the cloud server together to interactively generate a proof of possession of shared data. In ProofVerify, the public verifier audits the integrity of shared data by verifying the proof.

Note that for the ease of understanding, we first assume the group is static, which means the group is pre-defined before shared data is created in the cloud and the membership of the group is not changed during data sharing. Specifically, before the original user outsources shared data to the cloud, he/she decides all the group members. We will discuss the case of dynamic groups later.

Discussion. In the construction of Oruta, we support data privacy by leveraging random masking (i.e., $t_l h(Z_l)$ in ProofGen), which is also used in previous work [5] to protect data privacy for personal users. If a user wants to protect the content of private data in the cloud, this user can also encrypt data before outsourcing it into the cloud server with encryption techniques [30], [31], such as the combination of symmetric key encryption and attribute-based encryption (ABE) [30].

With the sampling strategy [9], which is widely used in most of the public auditing mechanisms, a public verifier can detect any corrupted block in shared data with a high probability by only choosing a subset of all blocks (i.e., choosing c-element subset I from set $[1; n]$) in each auditing task. Previous work [9] has already proved that, given a total number of blocks $n = 1{,}000{,}000$, if 1 percent of all the blocks are lost or removed, a public verifier can detect these corrupted blocks with a probability greater than 99 percent by choosing only 460 random blocks. Of course, this public verifier can always spend more communication overhead, and verify the integrity of data by choosing all the n blocks in shared data. Even if all the n blocks in shared data are selected (i.e., without using sampling strategy), the communication overhead during public auditing is still much more smaller than retrieving the entire data from the cloud [9].

Besides choosing a larger number of random blocks, another possible approach to improve the detection probability is to perform multiple auditing tasks on the same shared data by using different randoms (i.e., y_j is different for block m_j in each different task). Specifically, if the current detection probability is P_x and a number of t auditing tasks is performed, then the detection probability is computed as $1 - (1 - P_x)^t$.

Dynamic Groups. We now discuss the scenario of dynamic groups under our proposed mechanism. If a new user can be added in the group or an existing user can be revoked from the group, then this group is denoted as a dynamic group. To support dynamic groups while still allowing the public verifier to perform public auditing, all the ring signatures on shared data need to be re-computed with the signer's private key and all the current users' public keys when the membership of the group is changed.

For example, if the current size of the group is d and a new user u_{d+1} is added into the group, then a ring signature on each block in shared data needs to be re-computed with the signer's private key and all the $d + 1$ public keys $(pk_1; \ldots ; pk_{d+1})$. If the current size of the group is d and an existing user u_d is revoked from the group, then a ring signature on each block in shared data needs to be re-computed with the signer's private key and all the $d - 1$ public keys $(pk_1; \ldots ; pk_{d-1})$.

The main reason of this type of re-computation on signatures introduced by dynamic groups, is because the generation of a ring signature under our mechanism requires the signer's private key and all the current members' public keys. An interesting problem for our future work will be how to avoid this type of re-computation introduced by dynamic groups while still preserving identity privacy from the public verifier during the process of public auditing on shared data.

Security Analysis of Oruta

Now, we discuss security properties of Oruta, including its correctness, unforgeability, identity privacy and data privacy.

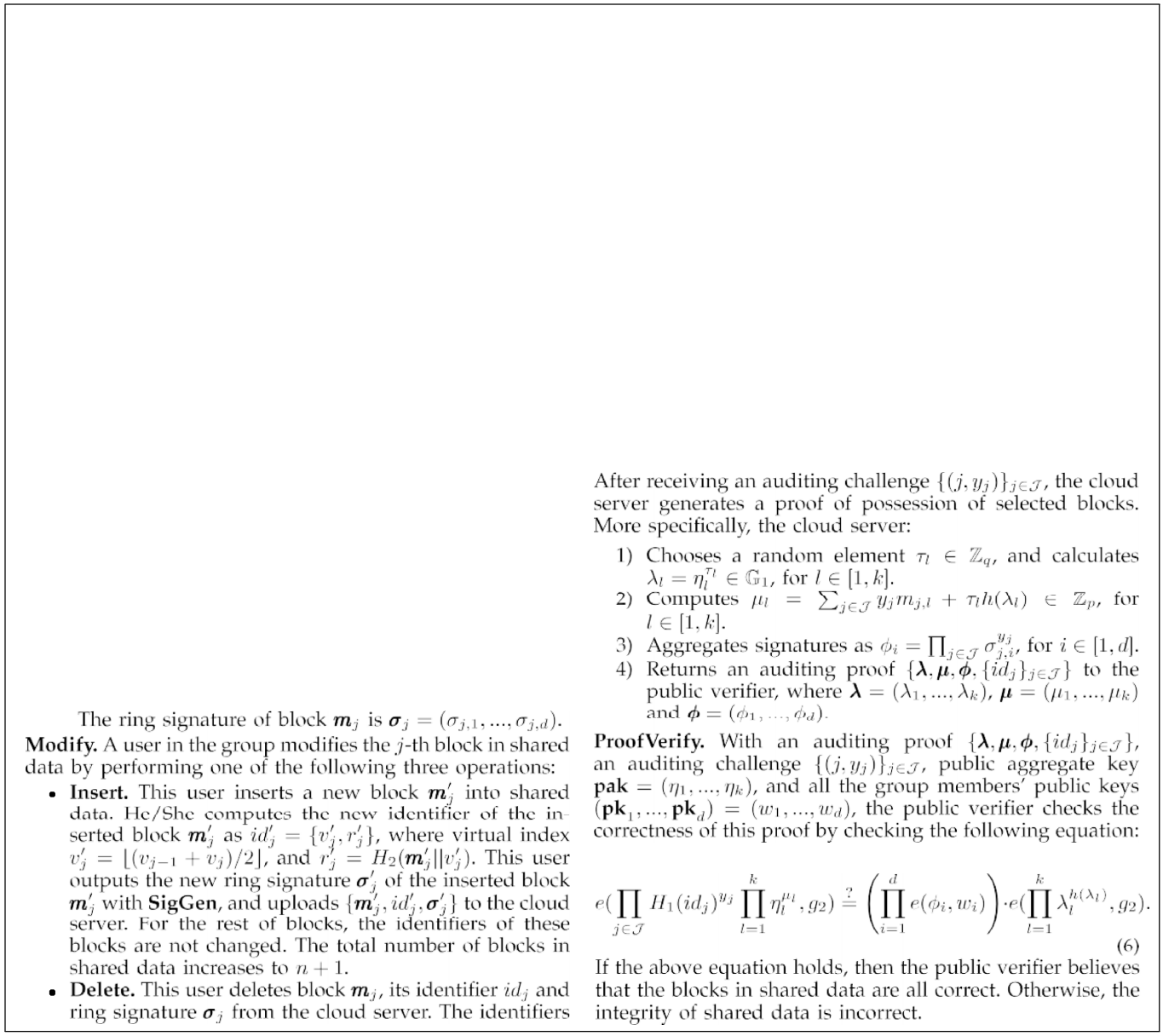
Theorem 5. *A public verifier is able to correctly audit the integrity of shared data under Oruta.*

Proof. According to the description of ProofVerify, a public verifier believes the integrity of shared data is correct if Equation (6) holds. So, the correctness of our scheme can be proved by verifying the correctness of Equation (6). Based on properties of bilinear maps and Theorem 1, the right-hand side (RHS) of Equation (6) can be expanded as follows:

$$
\begin{aligned}
\text{RHS} &= \prod_{i=1}^{d} e\Big(\prod_{j \in J} s_{j,i}^{y_j}, w_i\Big) \cdot e\Big(\prod_{l=1}^{k} Z^{h(t_l)}, g_2\Big) \\
&= \prod_{j \in J} \prod_{i=1}^{d} e(s_{j,i}^{y_j}, w_i) \cdot e\Big(\prod_{l=1}^{k} h_l^{t_l h(Z_l)}, g_2\Big) \\
&= \prod_{j \in J} e(b^j, g_2)^{y_j} \cdot e\Big(\prod_{l=1}^{k} h^{t_l h(Z_l)}, g_2\Big) \\
&= e\Big(\prod_{j \in J} H(id)^{y_j} \prod_{l=1}^{k} h^{m_{j,l}}, g\Big) \cdot e\Big(\prod_{l=1}^{k} h^{t_l h(Z_l)}, g\Big) \\
&= e\Big(\prod_{j \in J} \prod_{l=1}^{k} h^{y_j m_{j,l}}, g\Big) \cdot e\Big(\prod_{l=1}^{k} h^{t_l h(Z_l)}, g\Big) \\
&= e\Big(\prod_{j \in J} H(id)^{y_j} \cdot \prod_{l=1}^{k} h^{\sum_{j \in J} y_j m_{j,l} + t_l h(Z_l)}, g\Big) \\
&= \prod_{j \in J} H_1(id_j)^{y_j} \cdot \prod_{l=1}^{k} h_l^{m_l}; g_2 :
\end{aligned}
$$

Theorem 6. *For an untrusted cloud, it is computationally infeasible to generate a forgery of an auditing proof under Oruta as long as the DL assumption holds.*

ISBN 978-81-939386-1-4



The ring signature of block $\boldsymbol{m}_j$ is $\boldsymbol{\sigma}_j = (\sigma_{j,1}, ..., \sigma_{j,d})$.

Modify. A user in the group modifies the j-th block in shared data by performing one of the following three operations:

- **Insert.** This user inserts a new block $\boldsymbol{m}'_j$ into shared data. He/She computes the new identifier of the inserted block $\boldsymbol{m}'_j$ as $id'_j = \{v'_j, r'_j\}$, where virtual index $v'_j = \lfloor (v_{j-1} + v_j)/2 \rfloor$, and $r'_j = H_2(\boldsymbol{m}'_j || v'_j)$. This user outputs the new ring signature $\boldsymbol{\sigma}'_j$ of the inserted block $\boldsymbol{m}'_j$ with **SigGen**, and uploads $\{\boldsymbol{m}'_j, id'_j, \boldsymbol{\sigma}'_j\}$ to the cloud server. For the rest of blocks, the identifiers of these blocks are not changed. The total number of blocks in shared data increases to $n+1$.
- **Delete.** This user deletes block $\boldsymbol{m}_j$, its identifier id_j and ring signature $\boldsymbol{\sigma}_j$ from the cloud server. The identifiers

After receiving an auditing challenge $\{(j, y_j)\}_{j \in \mathcal{J}}$, the cloud server generates a proof of possession of selected blocks. More specifically, the cloud server:

1) Chooses a random element $\tau_l \in \mathbb{Z}_q$, and calculates $\lambda_l = \eta_l^{\tau_l} \in \mathbb{G}_1$, for $l \in [1, k]$.
2) Computes $\mu_l = \sum_{j \in \mathcal{J}} y_j m_{j,l} + \tau_l h(\lambda_l) \in \mathbb{Z}_p$, for $l \in [1, k]$.
3) Aggregates signatures as $\phi_i = \prod_{j \in \mathcal{J}} \sigma_{j,i}^{y_j}$, for $i \in [1, d]$.
4) Returns an auditing proof $\{\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\phi}, \{id_j\}_{j \in \mathcal{J}}\}$ to the public verifier, where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_k)$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)$ and $\boldsymbol{\phi} = (\phi_1, ..., \phi_d)$.

ProofVerify. With an auditing proof $\{\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\phi}, \{id_j\}_{j \in \mathcal{J}}\}$, an auditing challenge $\{(j, y_j)\}_{j \in \mathcal{J}}$, public aggregate key $\mathbf{pak} = (\eta_1, ..., \eta_k)$, and all the group members' public keys $(\mathbf{pk}_1, ..., \mathbf{pk}_d) = (w_1, ..., w_d)$, the public verifier checks the correctness of this proof by checking the following equation:

$$e(\prod_{j \in \mathcal{J}} H_1(id_j)^{y_j} \prod_{l=1}^{k} \eta_l^{\mu_l}, g_2) \stackrel{?}{=} \left(\prod_{i=1}^{d} e(\phi_i, w_i)\right) \cdot e(\prod_{l=1}^{k} \lambda_l^{h(\lambda_l)}, g_2). \quad (6)$$

If the above equation holds, then the public verifier believes that the blocks in shared data are all correct. Otherwise, the integrity of shared data is incorrect.

Fig. 7. Details of Oruta.

Proof. As proved in Theorem 2, for an untrusted cloud, if the Co-CDH problem on ðG1; G2Þ is hard, it is computationally infeasible to generate a forgery of a ring signature under HARS. In Oruta, besides trying to compute a forgery of a ring signature on each block to generate a forgery of an auditing proof, if the untrusted cloud could win the following security game, denoted as *Game 1*, it can generate a forgery of an auditing proof for corrupted shared data. Following the security game in [10], we describe this security game as follows:

Game 1. A public verifier sends an auditing challenge fð*j; yj*Þg*j*2J to the cloud server, the auditing proof generated based on the correct shared data *M* should be fZ;m;f;fid g g, which is able to pass the verification with Equation (6). The untrusted cloud generates a forg-

*j j*2J

*j*2J

nonzero since *M* 6 *M*. If this invalid proof based on the corrupted shared data *M* can successfully pass the verification, then the untrusted cloud wins. Otherwise, it fails.

Now, we prove that, if the untrusted cloud could win Game 1, we can find a solution of solving the DL problem in G1, which contradicts to the DL assumption that the DL problem in G1 is computationally infeasible. We first assume the untrusted cloud can win Game 1. Then, according to Equation (6), we have

e *H*1ð*id*jÞ*yj*· *k* *m*0 *l* *g*2 *l* 1 *j*2J *l* 1 *e*ð*f*i; *wi*Þ *e* *Z*hðZlÞ; *g*2 *:* *i*1 *l*1

also have

ISBN 978-81-939386-1-4

52 ery of the proof as $\{Z, \mathbf{m}', \mathbf{f}, \{id_j\}\}$ based on the corrupted shared data M, where $\mathbf{m} = (m_1, \ldots, m_k)$ and $\mathbf{m}' = (m'_1, \ldots, m'_k)$. Define $\Delta\mathbf{m} = (\Delta m_1, \ldots, \Delta m_k)$, where $\Delta m_l = m'_l - m_l$ for $1 \leq l \leq k$, and at least one element of $\Delta\mathbf{m}$ is non-zero.

Because $\{Z, \mathbf{m}', \mathbf{f}, \{id_j\}\}$ is a correct auditing proof, we

Let $\mathbb{G}_1, \mathbb{G}_2$ and $\mathbb{G}_T$ be multiplicative cyclic groups of order p, and g_1 is a generator of $\mathbb{G}_1$. Let g_2 be a generator of $\mathbb{G}_2$. Let $e : \mathbb{G}_1 \times \mathbb{G}_2 \to \mathbb{G}_T$ be a bilinear map, and $\psi : \mathbb{G}_2 \to \mathbb{G}_1$ be a computable isomorphism with $\psi(g_2) = g_1$. There are three hash functions H_1: $\{0,1\}^* \to \mathbb{G}_1$, $H_2 : \{0,1\}^* \to \mathbb{Z}_q$ and $h : \mathbb{G}_1 \to \mathbb{Z}_p$. The total number of users in the group is d. Shared data M is divided into n blocks as $M = (\boldsymbol{m}_1, ..., \boldsymbol{m}_n)$, and each block $\boldsymbol{m}_j = (m_{j,1}, ..., m_{j,k})$ is further divided into k elements of $\mathbb{Z}_p$. The global parameters are $(e, \psi, p, q, \mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T, g_1, g_2, H_1, H_2, h, d, n, k)$.

KeyGen. For user u_i, he/she randomly picks $x_i \in \mathbb{Z}_p$ and computes $w_i = g_2^{x_i}$. User u_i's public key is $\mathbf{pk}_i = w_i$ and his/her private key is $\mathbf{sk}_i = x_i$. The original user also randomly generates a public aggregate key $\mathbf{pak} = (\eta_1, ..., \eta_k)$, where η_l are random elements of $\mathbb{G}_1$.

SigGen. Given all the d group members' public keys $(\mathbf{pk}_1, ..., \mathbf{pk}_d) = (w_1, ..., w_d)$, a block $\boldsymbol{m}_j = (m_{j,1}, ..., m_{j,k})$, its identifier id_j, a private key $\mathbf{sk}_s$ for some s, user u_s computes a ring signature of this block as follows:

1) Aggregates block $\boldsymbol{m}_j$ with the public aggregate key $\mathbf{pak}$, and computes

$$\beta_j = H_1(id_j) \prod_{l=1}^{k} \eta_l^{m_{j,l}} \in \mathbb{G}_1. \tag{4}$$

2) Randomly chooses $a_{j,i} \in \mathbb{Z}_p$ and sets $\sigma_{j,i} = g_1^{a_{j,i}}$, for all $i \neq s$. Then, calculates

$$\sigma_{j,s} = \left(\frac{\beta_j}{\psi(\prod_{i \neq s} w_i^{a_{j,i}})} \right)^{1/x_s} \in \mathbb{G}_1. \tag{5}$$

- **Update.** This user updates the j-th block in shared data with a new block $\boldsymbol{m}'_j$. The virtual index of this block is remain the same, and r'_j is computed as $r'_j = H_2(\boldsymbol{m}'_j || v_j)$. The new identifier of this updated block is $id'_j = \{v_j, r'_j\}$. The identifiers of other blocks in shared data are not changed. This user outputs the new ring signature $\boldsymbol{\sigma}'_j$ of this new block with **SigGen**, and uploads $\{\boldsymbol{m}'_j, id'_j, \boldsymbol{\sigma}'_j\}$ to the cloud server. The total number of blocks in shared data is still n.

ProofGen. To audit the integrity of shared data, a public verifier:

1) Randomly picks a c-element subset $\mathcal{J}$ of set $[1, n]$ to locate the c selected blocks that will be checked, where n is total number of blocks in shared data.
2) Generates a random value $y_j \in \mathbb{Z}_q$, for $j \in \mathcal{J}$.
3) Sends an auditing challenge $\{(j, y_j)\}_{j \in \mathcal{J}}$ to the cloud server.

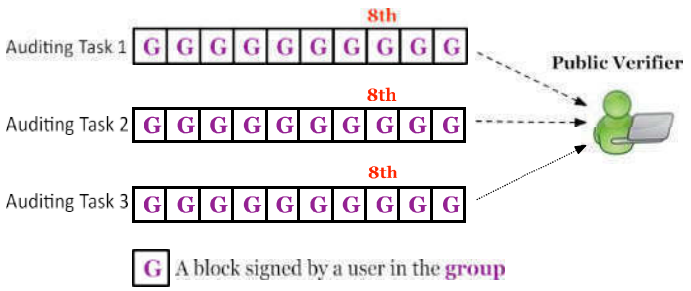


Fig. 8. Alice and Bob share a file in the cloud, and a public verifier audits the integrity of shared data with Oruta.

$$e\Big(\prod_{j\in J} H(id_j)^{y_j}\cdot \prod_{l=1}^{k} h_l^{m_l}, g_2\Big) \cdot e\Big(\prod_{i=1} e(f_i, w_i)^{h(Z_l)}, g_2\Big) = 1.$$

Then, we can learn that

$$\prod_{l=1} h_l^{m_l} = \prod_{l=1} h_l^{m_l}, \prod_{l=1} h_l^{m_l} = 1.$$

For two random elements $g, h \in G_1$, there exists $x \in Z_p$ that $h = g^x$ because G_1 is a cyclic group. Without loss of generality, given $g, h \in G_1$, each h_l is able to randomly and correctly generated by computing $h_l = g^{\alpha_l} h^{\beta_l} \in G_1$, where α_l and β_l are random values of Z_p. Then, we have

$$1 = \prod_{l=1} h_l^{m_l} = \prod_{l=1} g^{\alpha_l m_l} h^{\beta_l m_l} = g^{\sum_{l=1} \alpha_l m_l} \cdot h^{\sum_{l=1} \beta_l m_l}.$$

Clearly, we can find a solution to the DL problem. More specifically, given $g, h^x \in G_1$, we can compute

$$h = g^{\frac{\sum_{l=1} \alpha_l m_l}{\sum_{l=1} \beta_l m_l}} = g^x; \quad x = \frac{\sum_{l=1} \alpha_l m_l}{\sum_{l=1} \beta_l m_l}.$$

unless the denominator $\sum_{l=1} \beta_l m_l$ is zero. However, as we defined in Game 1, at least one element of $\{m_l \beta_l\}_{1 \le l \le k}$ is nonzero, and β_l is random element of Z_p, therefore, the denominator is zero with probability of $1/p$. It means, if the untrusted cloud wins Game 1, we can find a solution to the DL problem with a probability of $1 - 1/p$, which is non-negligible. It contradicts the assumption that the DL problem is hard in G_1. Therefore, for an untrusted cloud, it is computationally infeasible to generate a forgery of an auditing proof. □

Now, we show that the identity of the signer on each block in share data is not disclosed to a public verifier.

Theorem 7. *During public auditing, the probability for a public verifier to distinguish the identities of all the signers on the c selected blocks in shared data is at most $1/d^c$.*

Proof. According to Theorem 4, for any algorithm A, the probability to reveal the signer on one block is $1/d$. Because the c selected blocks are signed independently,

Assume the integrity of B auditing tasks need to be verified, where the B corresponding shared data are denoted as $M_1, ..., M_B$, the number of users sharing data M_b is described as d_b, where $1 \le b \le B$.

BatchProofGen. A public verifier first generates an auditing challenge $\{(j, y_j)\}_{j\in\mathcal{J}}$ as in **ProofGen**. After receiving the auditing challenge, the cloud server generates and returns an auditing proof $\{\boldsymbol{\lambda}_b, \boldsymbol{\mu}_b, \boldsymbol{\phi}_b, \{id_{b,j}\}_{j\in\mathcal{J}}\}$ for each shared data M_b as in **ProofGen**, where $1 \le b \le B$, $1 \le l \le k$, $1 \le i \le d_b$ and

$$\lambda_{b,l} = \eta_{b,l}^{\tau_{b,l}}$$
$$\mu_{b,l} = \sum_{j\in\mathcal{J}} y_j m_{b,j,l} + \tau_{b,l} h(\lambda_{b,l})$$
$$\phi_{b,i} = \sum_{j\in\mathcal{J}} \sigma_{b,j,i}^{y_j}$$

Here $id_{b,j}$ is described as $id_{b,j} = \{f_b, v_{b,j}, r_{b,j}\}$, where f_b is the data identifier (e.g., file name) of shared data M_b.

BatchProofVerify. After receiving all the B auditing proofs, the public verifier checks the correctness of these B proofs simultaneously by checking the following equation with all the $\sum_{b=1}^{B} d_b$ users' public keys:

$$e\Big(\prod_{b=1}^{B}\Big(\prod_{j\in\mathcal{J}} H(id_{b,j})^{y_j}\cdot \prod_{l=1}^{k} \eta_{b,l}^{\mu_{b,l}}\Big), g_2\Big) \stackrel{?}{=} \Big(\prod_{b=1}^{B}\prod_{i=1}^{d_b} e(\phi_{b,i}, w_{b,i})\Big)\cdot e\Big(\prod_{b=1}^{B}\prod_{l=1}^{k} \lambda_{b,l}^{h(\lambda_{b,l})}, g_2\Big), \quad (7)$$

where $\mathbf{pk}_{b,i} = w_{b,i} = g^{x_{b,i}}$ and $\mathbf{sk}_{b,i} = x_{b,i}$. If the above verification equation holds, then the public verifier believes that the integrity of all the B shared data is correct. Otherwise, there is at least one shared data is corrupted.

Fig. 9. Details of Batch Auditing.

where $c \in [1, n]$, the total probability that the public verifier can distinguish all the signers' identities on the c selected blocks in shared data is at most $1/d^c$. □

Let us reconsider the example we described in Section 1. With our proposed mechanism, the public verifier knows each block in shared data is either signed by Alice or Bob, because it needs both users' public keys to verify the correctness of the entire shared data. However, it cannot distinguish who is the signer on each particular block (as shown in Fig. 8). Therefore, the public verifier cannot have additional advantages on revealing private information, such as who always signs the largest number of blocks in shared data or which particular block is frequently modified by different group members.

Following the similar theorem in [5], we show that our scheme is also able to support data privacy.

Theorem 8. Given an auditing proof $\{Z, m, \sigma, \{id_j\}_{j\in J}\}$, it is computationally infeasible for a public verifier to reveal any private data in shared data under Oruta as long as the DL assumption holds.

Proof. If the combined element $\sum_{j\in J} y_j m_{j,l}$ which is a linear combination of all the elements in block m_j, is directly sent to a public verifier, the public verifier could learn the content of data by solving linear equations after collecting a sufficient number of linear combinations. To preserve private data, the combined element is computed with random masking as $m_l = \sum_{j\in J} y_j m_{j,l} + t_l h(Z_l)$. In

order to still solve linear equations, the public verifier must know the value of $t_l \in Z_p$. However, given $h_l \in G_1$ $Z_l = h_l^{t_l} \in G_1$, computing t_l is as hard as solving the DL problem in G_1, which is computationally infeasible. Therefore, give Z and $\mathbf{m}$, the public verifier cannot directly obtain any linear combination of all the elements in a block, and cannot further reveal any private data in shared data M.

Batch Auditing

Sometimes, a public verifier may need to verify the correctness of multiple auditing tasks in a very short time. Directly verifying these multiple auditing tasks separately would be inefficient. By leveraging the properties of bilinear maps, we can further extend Oruta to support batch auditing, which can verify the correctness of multiple auditing tasks simultaneously and improve the efficiency of public auditing. Details of batch auditing are presented in Fig. 9.

Based on the correctness of Equation (6), the correctness of batch auditing in Equation (7) can be presented as

$$\prod_{b=1}^{d_b} e(\mathbf{f}_{;w_{b;i}}) \cdot e\Big(\prod_{b=1}^{B} Z_{b;l}^{h(\mathbf{f}_{b;i})}, g_2\Big) = e\Big(\prod_{b=1}^{B}\prod_{i=1}^{k} \mathbf{f}_{;w}^{h(\mathbf{f}_{b;i})} \cdot e\Big(Z^{h(\mathbf{f}_{b;i})}, g_2\Big)\Big)$$
$$= e\Big(\prod_{b=1}^{B}\prod_{j\in J} H(id_{b;j})^{y_j} \cdot \prod_{l=1}^{k} h_{b;l}^{m_{b;l}}, g_2\Big)$$
$$= e\Big(\prod_{b=1}^{B}\prod_{j\in J} H(id_{b;j})^{y_j} \cdot \prod_{l=1}^{k} h_{b;l}^{m_{b;l}}, g_2\Big).$$

If all the B shared data are from the same group, the public verifier can further improve the efficiency of batch auditing by verifying

$$e\Big(\prod_{b=1}^{B}\prod_{j\in J} H(id_{b;j})^{y_j} \cdot \prod_{l=1}^{k} h_{b;l}^{m_{b;l}}, g_2\Big) \stackrel{?}{=} \prod_{i=1}^{d_b} e\Big(\mathbf{f}_{b;i;} w_i\Big) \cdot e\Big(\prod_{b=1}^{B}\prod_{l=1}^{k} Z_{b;l}^{h(\mathbf{f}_{b;l})}, g_2\Big), \quad (8)$$

which can save the public verifier about $(d - 1)B$ pairing operations in total compared to Equation (7).

Note that batch auditing will fail if at least one incorrect auditing proof exists in all the B auditing proofs. To allow most of auditing proofs to still pass the verification when there exists only a small number of incorrect auditing proofs, we can utilize binary search [5] during batch auditing.

More specifically, once the batch auditing of the B auditing proofs fails, the public verifier divides the set of all the B auditing proofs into two subsets, where each subset contains a number of B=2 auditing proofs. Then the public verifier re-checks the correctness of auditing proofs in each subset using batch auditing. If the verification result of one subset is correct, then all the auditing proofs in this subset are all correct. Otherwise, this subset is further divided into two sub-subsets, and the public verifier re-checks the correctness of auditing proofs in each sub-subset with batch auditing until all the incorrect auditing proofs are found. Clearly, when the number of incorrect auditing proofs increases, the public verifier needs more time to distinguish all the incorrect auditing proofs, and the efficiency of batch auditing will be reduced. Experimental results in Section 6 shows that, when less than 12 percent of all the B auditing proofs are incorrect, batching auditing is still more efficient than verifying all the B auditing proofs one by one.

6 Performance

In this section, we first analyze the computation and communication costs of Oruta, and then evaluate the performance of Oruta in experiments.

Computation Cost

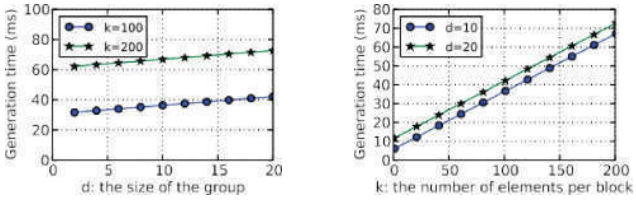
During an auditing task, the public verifier first generates some random values to construct an auditing challenge, which only introduces a small cost in computation. Then, after receiving the auditing challenge, the cloud server needs to compute an auditing proof $\{Z; \mathbf{m}; \mathbf{f}; \mathbf{f}_{id_j}\}_{j\in J}$. Based on the description in Section 5, the computation cost of calculating an auditing proof is about $(k + dc)\mathrm{Exp}_{G_1} + dc\mathrm{Mul}_{G_1} + ck\mathrm{Mul}_{Z_p} + k\mathrm{Hash}_{Z_p}$, where Exp_{G_1} denotes the cost of computing one exponentiation in G_1, Mul_{G_1} denotes the cost of computing one multiplication in G_1, Mul_{Z_p} and Hash_{Z_p} respectively denote the cost of computing one multiplication and one hashing operation in Z_p. To check the correctness of an auditing proof $\{Z; \mathbf{m}; \mathbf{f}; \mathbf{f}_{id_j}\}_{j\in J}$, a public verifier audits it with Equation (6). The total cost of verifying this auditing proof is about $(2k + d)\mathrm{Exp}_{G_1} + (2k + c)\mathrm{Mul}_{G_1} + d\mathrm{Mul}_{G_T} + c\mathrm{Hash}_{G_1} + (d + 2)\mathrm{Pair}$. We use Pair to denote the cost of computing one pairing operation on $e : G_1 \times G_2 \to G_T$.

Communication Cost

The communication cost of Oruta is mainly introduced by two aspects: the auditing challenge and auditing proof. For each auditing challenge $\{j, y_j\}_{j\in J}$, the communication cost is $c(|q| + |n|)$ bits, where $|q|$ is the length of an element of Z_q and $|n|$ is the length of an index. Each auditing proof $\{Z; \mathbf{m}; \mathbf{f}; \mathbf{f}_{id_j}\}_{j\in J}$ contains $(k + d)$ elements of G_1, k elements of Z_p and c elements of Z_q, therefore the communication cost of one auditing proof is $(2k + d)|p| + c|q|$ bits.

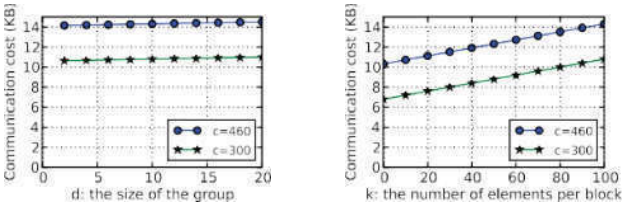
Experimental Results

We now evaluate the efficiency of Oruta in experiments. In our experiments, we utilize the GNU Multiple Precision Arithmetic (GMP) library and Pairing Based Cryptography (PBC) library. All the following experiments are based on C and tested on a 2.26 GHz Linux system over 1,000 times. Because Oruta needs more exponentiations than pairing operations during the process of auditing, the elliptic curve we choose in our experiments is an MNT curve with a base field size of 159 bits, which has a better performance than other curves on computing exponentiations. We choose $|p| = 160$ bits and $|q| = 80$ bits. We assume the total number of blocks in shared data is $n = 1{,}000{,}000$ and $|n| = 20$ bits. The size of shared data is 2 GB. To keep the detection probability greater than 99 percent, we set the number of selected



(a) Impact of d on signature generation time (ms). (b) Impact of k on signature generation time (ms).

Fig. 10. Performance of signature generation.



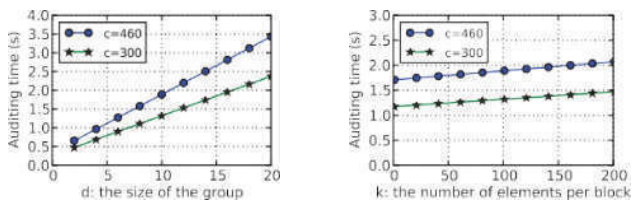
(a) Impact of d on communication cost (KB), where $k = 100$. (b) Impact of k on communication cost (KB), where $d = 10$.

Fig. 12. Performance of communication cost.

blocks in an auditing task as c ¼ 460 [9]. If only 300 blocks are selected, the detection probability is greater than 95 percent. We also assume the size of the group d 2 ½2; 20] in the following experiments. Certainly, if a larger group size is used, the total computation cost will increase due to the increasing number of exponentiations and pairing operations.

Performance of Signature Generation. According to Section 5, the generation time of a ring signature on a block is determined by the number of users in the group and the number of elements in each block. As illustrated in Figs. 10a and 10b, when k is fixed, the generation time of a ring signature is linearly increasing with the size of the group; when d is fixed, the generation time of a ring signature is linearly increasing with the number of elements in each block. Specifically, when d ¼ 10 and k ¼ 100, a user in the group requires about 37 milliseconds to compute a ring signature on a block in shared data.

Performance of Auditing. Based on our proceeding analyses, the auditing performance of Oruta under different detection probabilities is illustrated in Figs. 11a and 12b, and Table 2. As shown in Fig. 11a, the auditing time is linearly increasing with the size of the group. When c ¼ 300, if there are two users sharing data in the cloud, the auditing time is only about 0:5 seconds; when the number of group member increases to 20, it takes about 2:5 seconds to finish the same auditing task. The communication cost of an auditing task under different parameters is presented in Figs. 12a and 12b. Compared to the size of entire shared data, the communication cost that a public verifier consumes in an auditing task is very small. It is clear in Table 2 that when maintaining a higher detection probability, a public verifier needs to consume more computation and communication overhead to finish the auditing task. Specifically, when c ¼ 300, it takes a public verifier 1:32 seconds to audit the correctness of shared data, where the size of shared data is 2 GB; when c ¼ 460, a public verifier needs 1:94 seconds to verify the integrity of the same shared data.



(a) Impact of d on auditing time (second), where $k = 100$. (b) Impact of k on auditing time (second), where $d = 10$.

Fig. 11. Performance of auditing time.

As we discussed in the previous section, the privacy performance of our mechanism depends on the number of members in the group. Given a block in shared data, the probability that a public verifier fails to reveal the identity of the signer is 1 — 1=d, where $d \leq 2$. Clearly, when the number of group members is larger, our mechanism has a better performance in terms of privacy. As we can see from Fig. 13a, this privacy performance increases with an increase of the size of the group.

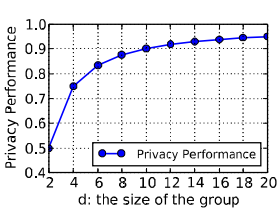
Performance of Batch Auditing. As we discussed in Section 5, when there are multiple auditing proofs, the public verifier can improve the efficiency of verification by performing batch auditing. In the following experiments, we choose c ¼ 300, k ¼ 100 and d ¼10. Compared to verifying a number of B auditing proofs one by one, if these B auditing proofs are for different groups, batching auditing can save 2:1 percent of the auditing time per auditing proof on average (as shown in Fig. 14a). If these B auditing tasks are for the same group, batching auditing can save 12:6 percent of the average auditing time per auditing proof (as shown in Fig. 14b).

Now we evaluate the performance of batch auditing when incorrect auditing proofs exist among the B auditing proofs. As we mentioned in Section 5, we can use binary search in batch auditing, so that we can distinguish the incorrect ones from the B auditing proofs. However, the increasing number of incorrect auditing proofs will reduce the efficiency of batch auditing. It is important for us to find out the maximal number of incorrect auditing proofs exist in the B auditing proofs, where the batch auditing is still more efficient than separate auditing.

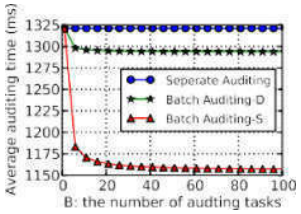
In this experiment, we assume the total number of auditing proofs in the batch auditing is B ¼ 128 (because we leverage binary search, it is better to set B as a power of 2), the number of elements in each block is k ¼ 100 and the number of users in the group is d ¼ 10. Let A denote the number of incorrect auditing proofs. In addition, we also assume that it always requires the *worst-case* algorithm to detect the incorrect auditing proofs in the experiment. According to Equation (7) and (8), the extra computation cost in binary search is mainly

TABLE 2
Performance of Auditing

System Parameters	$k = 100, d = 10,$	
Storage Usage	2GB + 200MB (data + signatures)	
Selected Blocks c	460	300
Communication Cost	14.55KB	10.95KB
Auditing Time	1.94s	1.32s

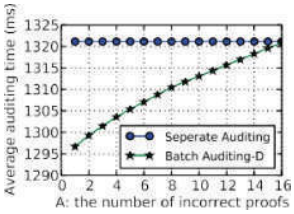


(a) Impact of d on privacy performance.

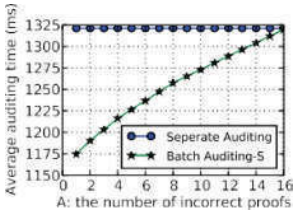


(b) Impact of B on the efficiency of batch auditing, where $k = 100$ and $d = 10$.

Fig. 13. Performance of privacy and batch auditing.



(a) Impact of A on the efficiency of batch auditing, where $B = 128$.



(b) Impact of A on the efficiency of batch auditing, where $B = 128$.

Fig. 14. Efficiency of batch auditing with incorrect proofs.

introduced by extra pairing operations. As shown in Fig. 14a, if all the 128 auditing proofs are for different groups, when the number of incorrect auditing proofs is less than 16 (12 percent of all the auditing proofs), batching auditing is still more efficient than separate auditing. Similarly, in Fig. 14b, if all the auditing proofs are for the same group, when the number of incorrect auditing proofs is more than 16, batching auditing is less efficient than verifying these auditing proofs separately.

7 Related Work

Provable data possession (PDP), proposed by Ateniese et al. [9], allows a verifier to check the correctness of a client's data stored at an untrusted server. By utilizing RSA-based homomorphic authenticators and sampling strategies, the verifier is able to publicly audit the integrity of data without retrieving the entire data, which is referred to as public auditing. Unfortunately, their mechanism is only suitable for auditing the integrity of personal data. Juels and Kaliski [32] defined another similar model called Proofs of Retrievability (POR), which is also able to check the correctness of data on an untrusted server. The original file is added with a set of randomly-valued check blocks called *sentinels*. The verifier challenges the untrusted server by specifying the positions of a collection of sentinels and asking the untrusted server to return the associated sentinel values. Shacham and Waters [10] designed two improved schemes. The first scheme is built from BLS signatures [27], and the second one is based on pseudo-random functions.

To support dynamic data, Ateniese et al. [33] presented an efficient PDP mechanism based on symmetric keys. This mechanism can support update and delete operations on data, however, insert operations are not available in this mechanism. Because it exploits symmetric keys to verify the integrity of data, it is not public verifiable and only provides a user with a limited number of verification requests. Wang et al. [12] utilized Merkle Hash Tree and BLS signatures [27] to support dynamic data in a public auditing mechanism. Erway et al. [11] introduced dynamic provable data possession (DPDP) by using authenticated dictionaries, which are based on rank information. Zhu et al. [15] exploited the fragment structure to reduce the storage of signatures in their public auditing mechanism. In addition, they also used index hash tables to provide dynamic operations on data. The public mechanism proposed by Wang et al. [5] and its journal version [18] are able to preserve users' confidential data from a public verifier by using random maskings. In addition, to operate multiple auditing tasks from different users efficiently, they extended their mechanism to enable batch auditing by leveraging aggregate signatures [21].

Wang et al. [13] leveraged homomorphic tokens to ensure the correctness of erasure codes-based data distributed on multiple servers. This mechanism is able not only to support dynamic data, but also to identify misbehaved servers. To minimize communication overhead in the phase of data repair, Chen et al. [14] also introduced a mechanism for auditing the correctness of data under the multi-server scenario, where these data are encoded by network coding instead of using erasure codes. More recently, Cao et al. [16] constructed an LT codes-based secure and reliable cloud storage mechanism. Compare to previous work [13], [14], this mechanism can avoid high decoding computation cost for data users and save computation resource for online data owners during datarepair.

8 Conclusion and Future Work

In this paper, we propose Oruta, a privacy-preserving public auditing mechanism for shared data in the cloud. We utilize ring signatures to construct homomorphic authenticators, so that a public verifier is able to audit shared data integrity without retrieving the entire data, yet it cannot distinguish who is the signer on each block. To improve the efficiency of verifying multiple auditing tasks, we further extend our mechanism to support batch auditing.

There are two interesting problems we will continue to study for our future work. One of them is traceability, which means the ability for the group manager (i.e., the original user) to reveal the identity of the signer based on verification metadata in some special situations. Since Oruta is based on ring signatures, where the identity of the signer is unconditionally protected [21], the current design of ours does not support traceability. To the best of our knowledge, designing an efficient public auditing mechanism with the capabilities of preserving identity privacy and supporting traceability is still open. Another problem for our future work is how to prove data freshness (prove the cloud possesses the latest version of shared data) while still preserving identity privacy.

Acknowledgments

This work was supported by NSFC 61272457, National Project 2012ZX03002003-002, 863 Project 2012AA013102, 111 Project B08038, IRT1078, FRF K50511010001 and NSFC 61170251. Most part of this work was done at University of

Toronto. A short version [1] of this paper is in Proceedings of the Fifth IEEE International Conference on Cloud Computing (IEEE Cloud 2012).

REFERENCES

[1] B. Wang, B. Li, and H. Li, "Oruta: Privacy-Preserving Public Auditing for Shared Data in the Cloud," *Proc. IEEE Fifth Int'l Conf. Cloud Computing*, pp. 295-302, 2012.

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing," *Comm. ACM*, vol. 53, no. 4, pp. 50-58, Apr. 2010.

[3] K. Ren, C. Wang, and Q. Wang, "Security Challenges for the Public Cloud," *IEEE Internet Computing*, vol. 16, no. 1, pp. 69-73, 2012.

[4] D. Song, E. Shi, I. Fischer, and U. Shankar, "Cloud Data Protection for the Masses," *Computer*, vol. 45, no. 1, pp. 39-45, 2012.

[5] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing," *Proc. IEEE INFOCOM*, pp. 525-533, 2010.

[6] B. Wang, M. Li, S.S. Chow, and H. Li, "Computing Encrypted Cloud Data Efficiently under Multiple Keys," *Proc. IEEE Conf. Comm. and Network Security (CNS '13)*, pp. 90-99, 2013.

[7] R. Rivest, A. Shamir, and L. Adleman, "A Method for Obtaining Digital Signatures and Public Key Cryptosystems," *Comm. ACM*, vol. 21, no. 2, pp. 120-126, 1978.

[8] The MD5 Message-Digest Algorithm (RFC1321). https://tools.ietf.org/html/rfc1321, 2014.

[9] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable Data Possession at Untrusted Stores," *Proc. 14th ACM Conf. Computer and Comm. Security (CCS '07)*, pp. 598-610, 2007.

[10] H. Shacham and B. Waters, "Compact Proofs of Retrievability," *Proc. 14th Int'l Conf. Theory and Application of Cryptology and Information Security: Advances in Cryptology (ASIACRYPT '08)*, pp. 90-107, 2008.

[11] C. Erway, A. Kupcu, C. Papamanthou, and R. Tamassia, "Dynamic Provable Data Possession," *Proc. 16th ACM Conf. Computer and Comm. Security (CCS'09)*, pp. 213-222, 2009.

[12] Q. Wang, C. Wang, J. Li, K. Ren, and W. Lou, "Enabling Public Verifiability and Data Dynamic for Storage Security in Cloud Computing," *Proc. 14th European Conf. Research in Computer Security (ESORICS'09)*, pp. 355-370, 2009.

[13] C. Wang, Q. Wang, K. Ren, and W. Lou, "Ensuring Data Storage Security in Cloud Computing," *Proc. 17th Int'l Workshop Quality of Service (IWQoS'09)*, pp. 1-9, 2009.

[14] B. Chen, R. Curtmola, G. Ateniese, and R. Burns, "Remote Data Checking for Network Coding-Based Distributed Storage Systems," *Proc. ACM Workshop Cloud Computing Security Workshop (CCSW'10)*, pp. 31-42, 2010.

[15] Y. Zhu, H. Wang, Z. Hu, G.-J. Ahn, H. Hu, and S.S Yau, "Dynamic Audit Services for Integrity Verification of Outsourced Storages in Clouds," *Proc. ACM Symp. Applied Computing (SAC'11)*, pp. 1550-1557, 2011.

[16] N. Cao, S. Yu, Z. Yang, W. Lou, and Y.T. Hou, "LT Codes-Based Secure and Reliable Cloud Storage Service," *Proc. IEEE INFOCOM*, 2012.

[17] B. Wang, B. Li, and H. Li, "Certificateless Public Auditing for Data Integrity in the Cloud," *Proc. IEEE Conf. Comm. and Network Security (CNS'13)*, pp. 276-284, 2013.

[18] C. Wang, S.S. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Secure Cloud Storage," *IEEE Trans. Computers*, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[19] B. Wang, B. Li, and H. Li, "Public Auditing for Shared Data with Efficient User Revocation in the Cloud," *Proc. IEEE INFOCOM*, pp. 2904-2912, 2013.

[20] B. Wang, B. Li, and H. Li, "Panda: Public Auditing for Shared Data with Efficient User Revocation in the Cloud," *IEEE Trans. Services Computing*, 20 Dec. 2013, DOI: 10.1109/TSC.2013.2295611.

[21] D. Boneh, C. Gentry, B. Lynn, and H. Shacham, "Aggregate and Verifiably Encrypted Signatures from Bilinear Maps," *Proc. 22nd Int'l Conf. Theory and Applications of Cryptographic Techniques: Advances in Cryptology (EUROCRYPT'03)*, pp. 416-432, 2003.

[22] B. Wang, H. Li, and M. Li, "Privacy-Preserving Public Auditing for Shared Cloud Data Supporting Group Dynamics," *Proc. IEEE Int'l Conf. Comm. (ICC'13)*, pp. 539-543, 2013.

[23] B. Wang, S.S. Chow, M. Li, and H. Li, "Storing Shared Data on the Cloud via Security-Mediator," *Proc. IEEE 33rd Int'l Conf. Distributed Computing Systems (ICDCS'13)*, pp. 124-133, 2013.

[24] D. Boneh, X. Boyen, and H. Shacham, "Short Group Signatures," *Proc. 24th Ann. Int'l Cryptology Conf. (CRYPTO'04)*, pp. 41-55, 2004.

[25] B. Wang, B. Li, and H. Li, "Knox: Privacy-Preserving Auditing for Shared Data with Large Groups in the Cloud," *Proc. 10th Int'l Conf. Applied Cryptography and Network Security (ACNS'12)*, pp. 507-525, June 2012.

[26] E. Brickell, J. Camenisch, and L. Chen, "Direct Anonymous Attestation," *Proc. 11th ACM Conf. Computer and Comm. Security (CCS'04)*, pp. 132-145, 2004.

[27] D. Boneh, B. Lynn, and H. Shacham, "Short Signatures from the Weil Pairing," *Proc. Seventh Int'l Conf. Theory and Application of Cryptology and Information Security: Advances in Cryptology (ASIACRYPT'01)*, pp. 514-532, 2001.

[28] R.L. Rivest, A. Shamir, and Y. Tauman, "How to Leak a Secret," *Proc. Seventh Int'l Conf. Theory and Application of Cryptology and Information Security: Advances in Cryptology (ASIACRYPT'01)*, pp. 552-565, 2001.

[29] D. Cash, A. Kupcu, and D. Wichs, "Dynamic Proofs of Retrievability via Oblivious RAM," *Proc. 32nd Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques: Advances in Cryptology (EUROCRYPT)*, pp. 279-295, 2013.

[30] S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving Secure, Scalable, and Fine-Grained Data Access Control in Cloud Computing," *Proc. IEEE INFOCOM*, pp. 534-542, 2010.

[31] X. Liu, Y. Zhang, B. Wang, and J. Yan, "Mona: Secure Multi-Owner Data Sharing for Dynamic Groups in the Cloud," *IEEE Trans. Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1182-1191, June 2013.

[32] A. Juels and B.S. Kaliski, "PORs: Proofs of Retrievability for Large Files," *Proc. 14th ACM Conf. Computer and Comm. Security (CCS'07)*, pp. 584-597, 2007.

[33] G. Ateniese, R.D. Pietro, L.V. Mancini, and G. Tsudik, "Scalable and Efficient Provable Data Possession," *Proc. Fourth Int'l Conf. Security and Privacy in Comm. Networks (SecureComm'08)*, 2008.

ISBN 978-81-939386-1-4

Ontology Frame Work Based On Generalized Rough Set Theory

A.prathap
Assistant Professor
Alaprathap055@gmail.com

Sudha pavani .K
Assistant Professor
sudhapavani1@gmail.com

K.Venkateswarlu
Assistant Professor
bagivenky@gmail.com

Abstract: -Ontology is a formal, explicit specification of a shared conceptualization. Ontologies are widely used in various domains such as medical, e-commerce and semantic web. Traditionally, ontology has been defined as the philosophical study of what exists, the study of kinds of entities in reality, and the relationships that these entities bear to one another. There are five types in ontologies namely Generic ontology, core ontology, domain ontology, task, and application ontology. Several mathematical techniques like fuzzy sets, graph theoretic methods are being used to define ontologies. Recently rough set theory has been used to develop ontologies. It extends the uncertainty representation in domain ontology and support uncertainty reasoning. The concepts of rough sets are being used to define ontologies on data with uncertainty. However, equivalence classes are used as building blocks. In addition to that implement an attribute dependency , how it impact the classification , and also analyze the reduct and core based on attribute dependencies Where concepts are defined using lower approximation .

Keywords: ontology, rough set theory, granular computing, attribute dependency, reduct and core

1. Introduction:

1.1 Ontology: Ontology is a formal, clear-cut description of a shared conceptualization. Ontologies are broadly utilized in different areas like medicine, e-trade and semantic web. Traditionally, ontology has been defined as the philosophical study of what exists, the investigation of sorts of elements truly, furthermore, the connections that these substances tolerate to each other.

At present ontologies have been extensively used by Artificial Intelligence as an unambiguous specification of conceptualizations in various areas, such as conceptual modeling, information integration and semantic web.

1.2 Rough sets: " The tuple consists of two approximations those are lower and upper approximation which is referred as rough set", here rough set is the collection of crisp sets, among these one shows the upper approximation of target set and the another

ISBN 978-81-939386-1-4

one shows lower approximation of target set.

Information System Framework: Consider I= (U,A) as an information system, here 'U' represents a non empty set which consists of predetermine objects and 'A' represents a non-empty set which represents the predetermined set of attributes. Consider the equation a: $U \rightarrow V_p$ for every $p \in A$. Where 'V_p' represents the set values that attribute 'p' might take. The information tables that assigns the value 'p(x)' from 'V_P' to each attributes which are exist in the 'P' and object 'x' in the Universe. For each $B \subseteq A$, there exist an associated equivalence relation 'IND(B)'

Let $R = IND(B) = \{(X, Y) \in U_2 / \forall p \in B, p(X) = p(Y)\}$

Example : structure of equivalence class

Object	a_1	a_2	a_3	a_4	a_5
O_1	1	2	0	1	1
O_2	1	2	0	1	1
O_3	2	0	0	1	0
O_4	0	0	1	2	1
O_5	2	1	0	2	1
O_6	0	0	1	2	2
O_7	2	0	0	1	0
O_8	0	1	2	2	1
O_9	2	1	0	2	2
O_{10}	2	0	0	1	0

Table 1.1: rough data set

When considering the set of attributes from $B = \{a_1, a_2, a_3, a_4, a_5\}$, it results following 7 equivalence classes.

$\{O_1, O_2\}$

$\{O_3, O_7, O_{10}\}$

$\{O_4\}$

$\{O_5\}$

$\{O_6\}$

$\{O_8\}$

$\{O_9\}$

Lower approximation: The R-lower estimation, represents the combination of all the equivalence classes in '$[x]_R$' those are enclosed with the target set.

Upper approximation: The R-upper estimation represents the combination of all the equivalence classes in $[x]_R$ which are having non-empty joint with the target set .

Attribute Dependency: The significant features of the data study or data attainment is the finding of attribute dependency that represent the finding the factors that are unequivocally identified with different factors. When all is said in done, it is these solid relationships that will merit extra examination, and that will in the long run be useful in the predictive modeling.

The representation of dependency can be handled elegantly in the theory of rough set. Assume two types of attributes those are set 'R' and the set 'S' and find out which degree of dependency attains in those. Here each and every attribute set simulate the structure of equivalence class. Consider the equivalence class that is stimulated by the set 'R' is represented by '$X_{[R]}$', and also the equivalence classes stimulated by set 'S' is denoted by'$X_{[S]}$'.

Consider $X_{[S]} = \{S_1, S_2, S_3, \ldots, S_N\}$ here, S_i represents the existing equivalence class from which the equivalence class structure is obtained by using set 'S' .

So, the attribute dependency of set 'S' on set 'R' is represented by $\gamma R(S)$.

$$\gamma R(S) = \frac{\sum_{i=1}^{N} |RS_i|}{|U|} \leq 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

Reduct and Core:

Reduct and core are two most Important ideas of unpleasant set hypothesis . Reduct is a diminished subset of unique set which holds the precision of unique set.

Reduct is frequently used in the attribute collection process to decrease un-necessary attributes. For example A is an original set, B is a sub set in A, where $B \subseteq A$, and $B_1, B_2, \ldots\ldots B_N$ are sub sets of B.

Where $U/IND(B_1) = U/IND(A)$

$U/IND(B_2) = U/IND(A)$

$U/IND(B_N) = U/IND(A)$

Then $B_1, B_2, \ldots\ldots B_N$ are the reducts and Core = $B_1 \quad B_2 \quad B_3 \quad \ldots\ldots\ldots \quad B_N$

In the example table above, attribute set $B_1 = \{a_3, a_4, a_5\}$ can be called as a reduct that means the attribute classes which comes from the set B_1 is equivalent to the original set 'A'.

$\{O_1, O_2\}$

$\{O_3, O_7, O_{10}\}$

$\{O_4\}$

$\{O_5\}$

$\{O_6\}$

$\{O_8\}$

$\{O_9\}$

From the above example the information system outcomes other reduct that is $B_2 = \{a1, a2, a5\}$, producing the equivalence class structure as same is represented with U / IND_A .

$\{O_1, O_2\}$

$\{O_3, O_7, O_{10}\}$

ISBN 978-81-939386-1-4

{O_4}

{O_5}

{O_6}

{O_8}

{O_9}

$IND(B_1)=IND(A)$

$IND(B_2)=IND(A)$

In the above example B_1 and B_2 are two reducts . The class of attributes which are frequent for each and every reduct is termed as a 'core'. So it consists of the attributes those cannot be altered from any other information system. The set of essential attributes will be considered by the core that means for the structure of the category which is going to represent. In the above example the type of attribute is {a_5}.

2. **Existing system:** In the existing system grainy calculations are used to construct ontologies. In the current framework for constructing ontologies utilized four diverse granular operations as a part of request to have various granular points of view for a particular ontological commitment. All things considered these four operations are utilized to have an assortment of perspectives of a metaphysic build with a rough set approach. So specifically rough set system is acquainted with build up a particular granular perspective of ontology. In this present reality recognizable proof of identicalness classes may not be conceivable, so with a specific end goal to maintain a strategic distance from here utilizing a similarity connection, infrequently likeness likewise called tolerance connection. In equivalence relation having "symmetric, transitive and reflexive relations". But in similarity only having "symmetric and reflexive", it simply slump the transitive condition.

3. **Proposed system:** In the existing system when the ε value is changed then the total classification is changed. The classification is sensitive to the value of ε , different values of ε produce different classification. So in the proposed system first analyses the representation of ontologies for all possible tolerances that can be generated for different values of ε . In addition to that implement an attribute dependency how it impact the classification, and also analyze the reduct and core based on attribute dependencies . Where concepts are defined using lower and upper approximation. Data set also created by using protégé tool for analysis.

4. Block diagram:

In this section first collect a data set ,after collecting data set now apply the

attribute dependency by using a lower approximation, based on attribute dependency reduct and core will be find. By using this core object build an ontology and classify it.

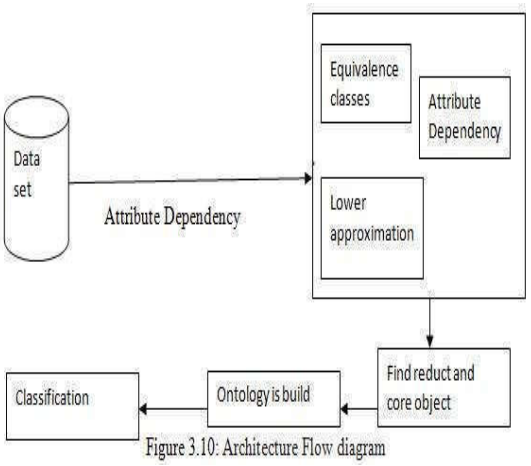


Figure 3.10: Architecture Flow diagram

5. Attribute dependency algorithm:

Input: Data set

Output: calculating attribute dependency

Method:

Let I=(U,A) . Having attributes $P=\{P_1,P_2,P_3\ldots\ldots P_N\}$ and objects are $Q=\{Q_1,Q_2,Q_3,\ldots\ldots Q_N\}$.

Step1: Calculate equivalence class family of 'IND(P)' and is represented by ' $U/IND(P)$

Step2: Calculate equivalence class family of 'IND(Q)' and is represented by $U/IND(Q)$.

Step3: Finding an lower approximation $\underline{P}Q$

Step4: Calculate attribute dependency $\gamma P(Q) = \frac{\sum |PQi|}{|U|} \leq 1$

END

6. Algorithm for Ontology building:

Input: Data set

Output: calculating attribute dependency

Method:

Let I=(U,A) Having two sets P and Q. $P=\{P_1,P_2,P_3\ldots\ldots P_N\}$ and $Q=\{Q_1,Q_2,Q_3,\ldots\ldots Q_N\}$

Step1: For i=1 to N

Step2: Remove P_i Calculate equivalence class family of 'IND(P)' and is represented by. $U/IND(P)$

Step3: Calculate equivalence class family of 'IND(Q)' and is represented by $U/IND(Q)$.

Step4: Finding an lower approximation $\underline{P}Q$

Step5: Calculate attribute dependency $\gamma P(Q) = \frac{\sum |PQi|}{|U|} \leq 1$

Step6: If($\gamma P(Q)$ == 1) remove P_i and P_{i+1} ,repeat 1,3 and 4 steps

Step7: Else , remove P_{i+1} and repeat 1,3 and 4 steps

Step8: Loop end.

Step9: calculate core and reduct on the data set based on Attribute dependency.

Step10: Build an ontology by using the out coming core object. That is level i(i=0…… n)

Step11: Remove the out coming core attribute from the data set ,and repeat the above 1 to 9 steps repeatedly .

Step12: calculate core until the remaining last attribute.

End

7. CONCLUSION

In this work a novel approach is adopted to build ontology based on the importance of the attributes at successive levels. Rough set framework has been adopted successfully and the concepts of core attributes has been applied. It can be observed that the ontology build using the proposed procedure is fixed and data dependent as against the existing methods in the literature which are either dependent on parameters or the ability of the designer.

The metrics to study the effective designs are varied the metrics designed for study of cohesion are studied. It is observed that the proposed method is producing meaningful and unique ontology structure.

8. References:

[1] Gruber, Tr. " A transalation approach to portable ontology specification"Knowledge Acquisition. 5: 199-220.

[2] Haining yao, Anthony Mark Orme and Letha Etzkon " Cohesion metrics for ontology design and application" journal of computer science 1(1) 10-113, 2005.

[3] Ian Horrocks, Peter F Patel Schneider, Frank van Harmelen "From SHIQ and RDF to OWL: the macking of a web ontology language" Science services and agents on the world wide web 1,7-26 (2003).

[4] Ian Horrocks, Peter patel Schneider" Reducing OWL entailement to description logic satisfiability" web semantics Science services and agents on the world wide web.1, 345-357(2004)

[5] John H. Genari. Mark A. Musen, Ray: W."FEergerson owl nentailment to description logic satisfiability" semantics web .Services and agents on the www58 (2003) 89-123

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Providing Security for Outsourced Data in Cloud Storage

MD.Riyaz
Assistant Professor , MRCE
Department of CSE
riyazjbs@gmail.com

A.Veerender
Assistant Professor , MRCE
Department of CSE
veerender57@gmail.com

CH.Malleswar Rao
Assistant Professor , MRCE
Department of CSE
malleswar.538@gmail.com

ABSTRACT —

To protect outsourced data in cloud storage against corruptions, adding fault tolerance to cloud storage together with data integrity checking and failure reparation becomes critical. Recently, regenerating codes have gained popularity due to their lower repair bandwidth while providing fault tolerance. Existing remote checking methods for regenerating-coded data only provide private auditing, requiring data owners to always stay online and handle auditing, as well as repairing, which is sometimes impractical. In this paper, we propose a public auditing scheme for the regenerating-code-based cloud storage. To solve the regeneration problem of failed authenticators in the absence of data owners, we introduce a proxy, which is privileged to regenerate the authenticators, into the traditional public auditing system model. Moreover, we
Design a novel public verifiable authenticator, which is generated by a couple of keys and can be regenerated using partial keys. Thus, our scheme can completely release data owners from online burden. In addition, we randomize the encode coefficients with a pseudorandom function to preserve data privacy. Extensive security analysis shows that our scheme is provable secure under random oracle model and experimental evaluation indicates that our scheme is highly efficient and can be feasibly integrated into the regenerating-code-based cloud storage.

Keywords- Cloud storage, regenerating codes, public audit, privacy preserving, authenticator regeneration, proxy, privileged, provable secure.

ISBN 978-81-939386-1-4

I. INTRODUCTION

CLOUD storage is gaining popularity as it gives a adaptable on-demand data outsourcing techniques services together with interesting advantages: comfort on the load regarding hard drive operations, general data accessibility together with location self-reliance, along with avoidance involving funds outlay upon components, software program, along with personalized maintenances, for example., [1]. On the other hand, this particular new paradigm involving data hosting services furthermore brings new safety measures provocations in the direction of users data, therefore making individuals or perhaps enterprisers nevertheless experience cautious. The most important work among most of these research tests would be the PDP (provable data possession) product and POR (proof of retrievability) product, which have been originally proposed for that single-server circumstance by Ateniese et 's. [2] and Juels et. 's. [3], respectively. Taking into consideration which documents usually are striped and redundantly located throughout multi-servers or even multi-clouds, [4]–[10] explore sincerity verification systems made for such multi-servers or even multiclouds environment with diverse redundancy systems, such as duplication, erasure code, and, lately, regenerating code. This paper concentrates on the particular integrity verification issue in regenerating-code-based cloud storage, particularly with the particular practical repair tactic [11]. Identical research studies have been performed through Bo Chen et 's. [7] as well as They would. Chen el 's. [7] [8] expanded the particular single-server CPOR scheme(private variation in [12]) towards the regenerating code-scenario; [8] created as well as applied a new data integrity protection(DIP) scheme regarding FMSR [13]-based cloud storage plus the scheme can be used towards the thin-cloud settings. For the huge size of the outsourced data as **well as the particular user's restricted source** functionality, the particular tasks of auditing as well as reparation in the cloud can be strong as well as expensive for the users [14]. The actual expense of making use of cloud storage should become minimized if you can , in ways that a new user won't have to execute a great number of operations for their outsourced data (in more in order to

locating it) [15]. Especially, people might not want to pass through the particular complexity in verifying as well as reparation. Your auditing plans in [7], [8] necessarily mean the challenge which people have to always remain on the web, which can slow down its usage in training, for long-term archival storage.

II. LITERATURE REVIEW

1. **Above the Clouds: A Berkeley View of Cloud Computing**

AUTHORS: Michael Armbrust

The long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service and shaping the way IT hardware is designed and purchased. Developers with innovative ideas for new Internet services no longer require the large capital outlays in hardware to deploy their service or the human expense to operate it. They need not be concerned about over- provisioning for a service whose popularity does not meet their predictions, thus wasting costly resources, or under-provisioning for one that becomes wildly popular, thus missing potential customers and revenue. Moreover, companies with large batch-oriented tasks can get results as quickly as their programs can scale, since using 1000 servers for one hour costs no more than using one server for 1000 hours. This elasticity of resources, without paying a premium for large scale, is unprecedented in the history of IT.

2. **PORs: Proofs of Retrievability for Large Files**

AUTHORS: Ari Juels Author defines and explore proofs of retrievability(PORs). A POR scheme enables an archive or back-up service (prover) to produce a concise proof that a user (verifier) can retrieve a target file F, that is, that the archive retains and reliably transmits file data sufficient for the user to recover F in its entirety. A POR may be viewed as a kind of cryptographic proof of knowledge (POK), but one specially designed to handle a large file (or bitstring) F. We explore POR protocols here in which the communication costs, number of memory accesses for the prover, and storage requirements of the user (verifier) are small

parameters essentially independent of the length of F. In addition to proposing new, practical POR constructions, author explores implementation considerations and optimizations that bear on previously explored, related schemes.

3. **MR-PDP: Multiple-Replica Provable Data Possession**

AUTHORS: Reza Curtmola Many storage systems rely on replication to increase the availability and durability of data on untrusted storage systems. At present, such storage systems provide no strong evidence that multiple copies of the data are actually stored. Storage servers can collude to make it look like they are storing many copies of the data, whereas in reality they only store a single copy. Authors address this short-coming through multiple-replica provable data possession (MR-PDP): A provably-secure scheme that allows a client that stores t replicas of a file in a storage system to verify through a challenge-response

4. **HAIL: A High-Availability and Integrity Layer for Cloud Storage**

AUTHORS: Kevin D. Bowers Author introduce HAIL (High-Availability and Integrity Layer), a distributed cryptographic system that permits a set of servers to prove to a client that a stored file is in- tact and retrievable. HAIL strengthens, formally unifies, and stream lines distinct approaches from the cryptographic and distributed-systems communities. Proofs in HAIL are efficiently computable by servers and highly compact typically tens or hundreds of bytes, irrespective of file size. HAIL cryptographically verifies and reactively reallocates file shares. It is robust against an active, mobile adversary, i.e., one that may progressively corrupt the full set of servers. Author proposes a strong, formal adversarial model for HAIL, and rigorous analysis and parameter choices. Author show how HAIL improves on the security and efficiency of existing tools, like Proofs of Retrievability(PORs)deployed on individual servers. Authors also report on a prototype implementation.

5. **Remote Data Checking for Network Coding-based Distributed Storage Systems**

AUTHORS: Bo Chen, Reza Curtmola

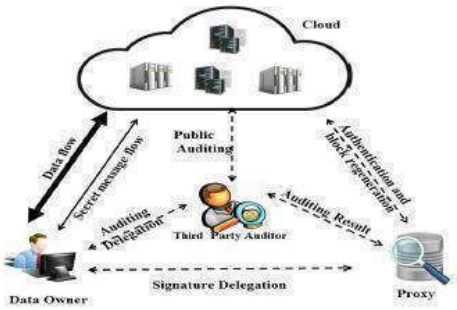
Remote Data Checking (RDC) is a technique by which clients can establish that data outsourced at untrusted servers remains intact over time. RDC is useful as a prevention tool, allowing clients to periodically check if data has been damaged, and as a repair tool whenever damage has been detected. Initially proposed in the context of a single server, RDC was later extended to verify data integrity in distributed storage systems that rely on replication and on erasure coding to store data redundantly at multiple servers. Recently, a technique was proposed to add redundancy based on network coding, which offers interesting tradeoffs because of its remarkably low communication overhead to repair corrupt servers. Management scheme generates an enormous number of keys with the increasing number users and requires users to dedicatedly protect the master keys. To this end, authors propose Dekey, a new construction in which users do not need to manage any keys on their own but instead securely distribute the convergent key shares across multiple servers. Security analysis demonstrates that Dekey is secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement Dekey using the Ramp secret sharing scheme and demonstrate that Dekey incurs limited overhead in realistic environments.

III. PROBLEM STATEMENT

Regenerating Codes: Regenerating codes are first introduced by A. G. Dimakis et al for distributed storage to reduce the repair bandwidth. Viewing cloud storage to be a collection of n storage servers, data file F is encoded and stored redundantly across these servers. Then F can be retrieved by connecting to any k-out-of-n servers, which is termed the MDS2-property. When data corruption at a server is detected, the client **will contact ℓ healthy servers and do**wnload **′ bits from each server, thus regenerating** the corrupted blocks without recovering the entire original file. Dimakis showed that the repair bandwidth can be significantly reduced with. Furthermore, they analyzed the fundamental tradeoff between the **storage cost α′ and the repair bandwidth ′,** then presented two extreme and practically

relevant points on the optimal tradeoff curve: the minimum bandwidth regenerating (MBR) point, which represents the operating point with the least possible repair bandwidth, and the minimum storage regenerating (MSR) point, which corresponds to the least possible storage cost on the servers. Denoted by the parameter tuple according to whether the corrupted blocks can be exactly regenerated, there are two versions of repair strategy: exact repair and functional repair. Exact repair strategy requires the repaired server to store an exact replica of the corrupted blocks, while functional repair indicates that the newly generated blocks are different from the corrupted ones with high probability. As one basis of our work, the functional repair regenerating codes are non-systematic and do not perform as well for read operation as systematic codes, but they really make sense for the scenario in which data repair occurs much more often than read, such as regulatory storage, data escrow and long-term archival storage.

IV. SYSTEM DESIGN



V. PROPOSED SYSTEM

In this paper we focus on the integrity verification problem in regenerating-code-based cloud storage, especially with the functional repair strategy. Similar studies have been performed by Bo Chen et al. and H. Chen el al. separately and independently.

Extend the single-server CPOR scheme(private version in) to the regenerating- code-scenario; designed and implemented a data integrity protection(DIP) scheme for FMSR based cloud storage and the scheme is adapted to the thin-cloud setting1. However, both of them are designed for private audit, only the data owner is allowed to verify the integrity and repair the faulty servers. Considering the large size of the outsourced data and the **user's constrained resource capability, the** tasks of auditing and reparation in the cloud

can be formidable and expensive for the users. The overhead of using cloud storage should be minimized as much as possible such that a user does not need to perform too many operations to their outsourced data (in additional to retrieving it)

VI. CONCLUSION

We propose a public auditing scheme for the regenerating-code-based cloud storage system, where the data owners are special to delegate TPA for their data validity checking. To secure the original data privacy against the TPA, we randomize the coefficients before all else rather than applying the visually impaired method amid the auditing procedure. Considering that the data owner cannot always stay online in practice, with a specific end goal to keep the storage available and verifiable after a malicious defilement, we bring a semi-trusted proxy into the system show and give a benefit to the proxy to handle the reparation of the coded pieces and authenticators. To better appropriate for the regenerating-code-scenario, we plan our authenticator based on the BLS signature. This authenticator can be efficiently generated by the data owner simultaneously with the encoding method. Broad analysis demonstrates that our scheme is provable secure, and the performance evaluation demonstrates that our scheme is profoundly efficient and can be feasibly integrated into a regenerating-code-based cloud storage.

VII. REFERENCES

[1] A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, **"Above the clouds: A Berkeley view of cloud computing,"** *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS*, vol. 28, p. 13, 2009.

[2] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. **Song, "Provable data possession at untrusted stores," in** *Proceedings of the 14th ACM Conference on Computer and Communications Security*, **ser. CCS '07.** New York, NY, USA: ACM, 2007, pp. 598–609.

[γ] **A. Juels and B. S. Kaliski Jr, "Pors:** Proofs of retrie**vability for large files," in** *Proceedings of the 14th ACM conference on Computer and communications security*.

ACM, 2007, pp. 584–597. [4] R. Curtmola, **O. Khan, R. Burns, and G. Ateniese, "Mr-**pdp: Multiplereplica provable data **possession," in** *Distributed Computing* ***Systems, 2008. ICDCS'08. The 28th*** *International Conference on*. IEEE, 2008, pp. 411–420.

[5] K. D. Bowers, A. Juels, and A. Oprea, **"Hail: a high-**availability and integrity layer **for cloud storage," in** *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 2009, pp. 187–198.

[6] J. He, Y. Zhang, G. Huang, Y. Shi, and **J. Cao, "Distributed data possession** checking for securing multiple replicas in **geographically dispersed clouds,"** *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1345–1358, 2012.

[7] B. Chen, R. Curtmola, G. Ateniese, and **R. Burns, "Remote data checking for** network coding-based distributed storage **systems," in** *Proceedings of the 2010 ACM workshop on Cloud computing security workshop*. ACM, 2010, pp. 31–42.

[8] H. Chen and P. Lee, "Enabling data integrity protection in regenerating coding-based cloud storage: Theory and **implementation,"** *Parallel and Distributed Systems, IEEE Transactions on*, vol. 25, no. 2, pp. 407–416, Feb 2014.

[9] K. Yang and X. Jia, "An efficient and secure dynamic auditing protocol for data **storage in cloud computing,"** *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 9, pp. 1717–1726, 2013.

[10] Y. Zhu, H. Hu, G.-J. Ahn, and M. Yu, **"Cooper**ative provable data possession for **integrity verification in multicloud storage,"** *Parallel and Distributed Systems, IEEE Transactions on*, vol. 23, no. 12, pp. 2231–2244, 2012.

[11] A. G. Dimakis, K. Ramchandran, Y. **Wu, and C. Suh, "A survey on network** code**s for distributed storage,"** *Proceedings of the IEEE*, vol. 99, no. 3, pp. 476–489, 2011.

[1β] H. Shacham and B. Waters, "Compact proofs of retrievability," in *Advances in Cryptology-ASIACRYPT 2008*. Springer, 2008, pp. 90– 107.

[13] Y. Hu, H. C. Chen, P. P. Lee, and Y. **Tang, "Nccloud: Applying network coding** for the storage repair in a cloud-of-**clouds,"** in *USENIX FAST*, 2012.

[14] C. Wang, Q. Wang, K. Ren, and W. **Lou, "Privacy**-preserving public auditing for **data storage security in cloud computing," in** *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[15] C. Wang, S. S. Chow, Q. Wang, K. **Ren, and W. Lou, "Privacy**-preserving **public auditing for secure cloud storage,"** *Computers, IEEE Transactions on*, vol. 62, no. 2, pp. 362–375, 2013.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Different Types of Machine Learning Techniques in Python programming

Mr Naveen Kumar R
M-Tech Student, CSE
Mallareddy College of Engineering
naveenit@gmail.com

Ms Adepu Aravinda
M-Tech Student, CSE
Mallareddy College of Engineering
aravinda1211@gmail.com

Mrs Subhashini Peneti
Assistant Professor ,CSE
Mallareddy College of Engineering
subhashinivalluru@gmail.com

***Abstract:* - When we hear "Machine learning", We will probably imagine a robot or something like the deadly Terminator as we see in movies. But, machine learning is involved not only in robotics[1], but also in many other applications. Now it is not just a futuristic fantasy, its already here. In fact, it has been around for decades in some specialized applications, such as Optical Character Recognition (OCR). But the first Machine Learning application that really became mainstream, improving the lives of hundreds of millions of people, took over the world back in the 1990s. It was the spam[2] filter. It does have all technical qualities to qualify as Machine Learning. It has, learned so well that you seldom need to flag an email as spam anymore. It was followed by hundreds of Machine Learning applications that now quietly power hundreds of products and features that you use regularly, from better recommendations to face recognition and voice search. This paper focuses on Explaining What is Machine Learning, why we need Machine Learning, Types of Machine Learning, Various Technologies available to implement Machine learning and how to use Python[3] as development environment for developing Machine Learning.**

Keywords:* *Machine Learning, training set, clustering, reinforcement, python, NumPy, SciPy, Scikit-learn, Python, pip, Spam, OCR, robotics.

I. INTRODUCTION

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people [1]. Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis to output values that fall will within a specific range of given output values. [4]

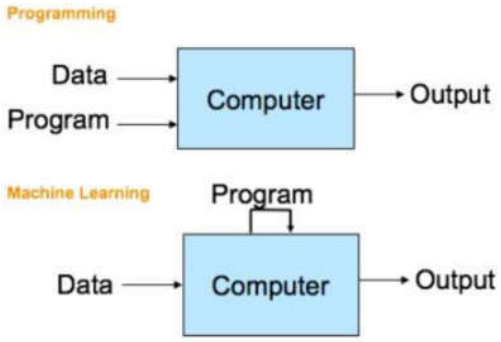


Fig. 1. Programming vs Machine Learning

In this age of modern technology, there is one resource that we have in abundance is large amount of structured and unstructured data. In the second half of the twentieth century, machine learning has evolved that involved the development of self-learning algorithms to gain knowledge from that data to make predictions. Instead of requiring humans to manually derive rules and build models from analysing large amounts of data, machine learning offers a more efficient alternative for capturing the knowledge in data to gradually improve the performance of predictive models and make data-driven decisions. Not only is machine learning becoming increasingly important in computer science research, but it also plays an ever-greater role in our everyday life. Thanks to machine learning, we enjoy robust e-mail spam filters, convenient text and voice recognition software, reliable Web search engines, challenging chess players, and, hopefully soon, safe and efficient self-driving cars.

[1] Technology dealing with design of robots
[2] Irrelevant or unwanted messages
[3] Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.

ISBN 978-81-939386-1-4

II. WHAT IS MACHINE LEARNING?

Machine Learning is the science of programming that gives computers the ability to learn without being explicitly programmed. Technically A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. [1]

For example, your spam filter is a Machine Learning program that can learn to flag spam given examples of spam emails (e.g., flagged by users) and examples of regular (non-spam or ham) emails. The examples that the system uses to learn are called the training set. Each training example is called a training instance (or sample). In this case, the task T is to flag spam for new emails, the experience E is the training data, and the performance measure P needs to be defined; for example, you can use the ratio of correctly classified emails. This, performance measure is called accuracy and it is often used in classification tasks.

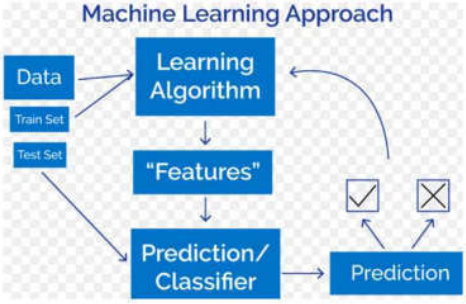


Fig. 2. Flow in Machine Learning

III. WHY USE MACHINE LEARNING?

Consider how to write a spam filter using traditional programming techniques.

1. First look at what spam typically looks like. We might notice that some words or phrases (such as "4U", "credit card", "free", and "amazing") tend to come up a lot in the subject. Perhaps we would also notice a few other patterns[4] in the sender's name, the email's body, and so on.[3]

2. We have to write a detection algorithm for each of the patterns that we noticed, and our program would flag emails as spam if a number of these patterns are detected.

3. We have to test the program and repeat steps 1 and 2 until it is good enough.

[4] Arrangement or words in a format

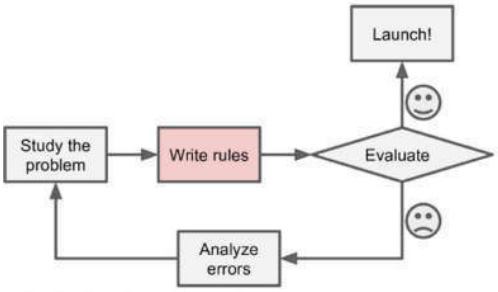


Fig. 3. Traditional approach

Since the problem is not trivial, the program will likely become a long list of complex rules which are hard to maintain.

In contrast, a spam filter based on Machine Learning techniques automatically learns which words and phrases are good predictors of spam by detecting unusually frequent patterns of words in the spam examples compared to the ham examples.

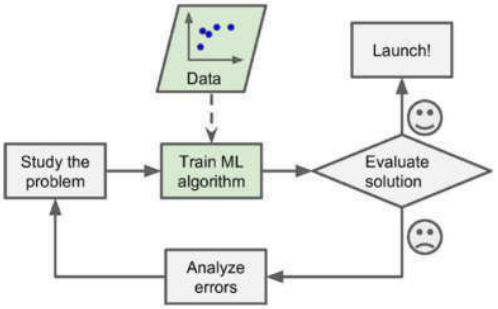


Fig. 4. Machine Learning approach

The program is much shorter, easier to maintain, and most likely more accurate. Moreover, if spammers notice that all their emails containing "4U" are blocked, they might start writing "For U" instead. A spam filter using traditional programming techniques would need to be updated to flag "For U" emails. If spammers keep working around the spam filter, we will need to keep writing new rules forever. In contrast, a spam filter based on Machine Learning techniques automatically notices that "For U" has become unusually frequent in spam flagged by users, and it starts flagging them without our intervention.

Another area where Machine Learning shines is for problems that either are too complex for traditional approaches or have no known algorithm. For example, consider speech recognition: say we want to start simple and write a program capable of distinguishing the words "one" and "two." You might notice that the word "two" starts with a high-pitch sound ("T"), so you could hardcode an algorithm that measures high-pitch sound intensity

and use that to distinguish ones and twos. Obviously, this technique will not scale to thousands of words spoken by millions of very different people in noisy environments and in dozens of languages. The best solution (at least today) is to write an algorithm that learns by itself, given many example recordings for each word.

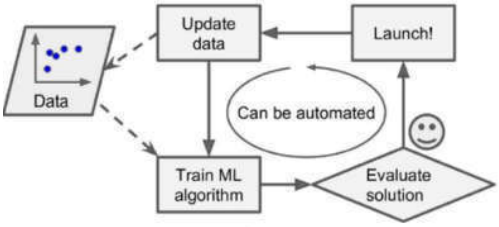


Fig. 5. Automatically adopting to change

To summarize, Machine Learning is great for:

- Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform better.
- Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.
- Fluctuating environments: a Machine Learning system can adapt to new data.
- Getting insights about complex problems and large amounts of data.

IV. HOW DOES MACHINE LEARNING WORK ?

Machine Learning algorithm is trained using a training data set to create a model. When new input data is introduced to the Machine Learning algorithm, it makes a prediction on the basis of the model. The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning algorithm is deployed. If the accuracy is not acceptable, the Machine Learning algorithm is trained again and again with an augmented training data set. This is just a very high-level example as there are many factors and other steps involved.[3]

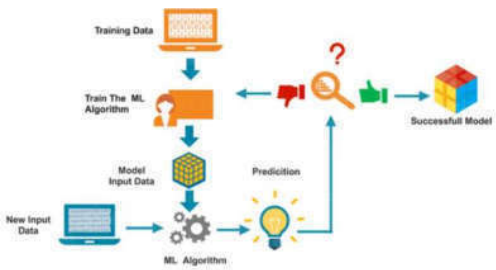


Fig. 6. Machine Learning Flowchart

V. TYPES OF MACHINE LEARNING

There are so many different types of Machine Learning systems that it is useful to classify them in broad categories. In this paper, we will look at the three types of machine learning: [4]

- Supervised learning,
- Unsupervised learning, and
- Reinforcement learning.

We will learn about the fundamental differences between the three different learning types and, using conceptual examples, we will develop an intuition for the practical problem domains where these can be applied.

A. Supervised Learning:

Supervised Learning is the one, where you can consider the learning is guided by a teacher. We have a dataset which acts as a teacher and its role is to train the model or the machine. Once the model gets trained it can start making prediction or decision when new data is given to it. In supervised learning, the training data you feed to the algorithm includes the desired solutions, called labels.[4]

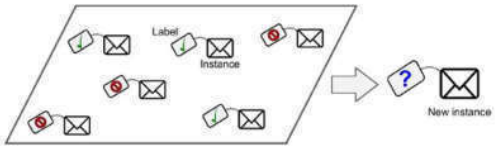


Fig. 7. Supervised Learning - Training set

A typical supervised learning task is classification. The spam filter is a good example of this: it is trained with many example emails along with their class (spam or ham), and it must learn how to classify new emails. Another typical task is to predict a target numeric value, such as the price of a car, given a set of features (mileage, age, brand, etc.) called predictors. To train the system, you need to give it many examples of cars, including both their predictors and their labels (i.e., their prices).

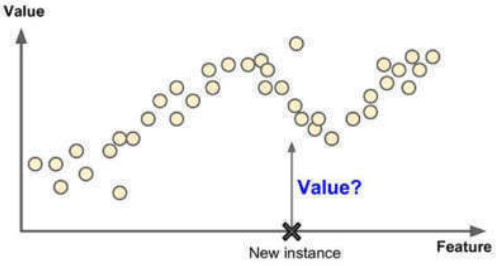


Fig. 8. Supervised Learning (Feature - Value graph)

Here are some of the most important

- k-Nearest Neighbours
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests
- Neural networks

B. Unsupervised Learning

The model learns through observation and finds structures in the data. Once the model is given a dataset, it automatically finds patterns and relationships in the dataset by creating clusters[5] in it. What it cannot do is add labels to the cluster, like it cannot say this a group of apples or mangoes, but it will separate all the apples from mangoes.[4]

Suppose we presented images of apples, bananas and mangoes to the model, so what it does, based on some patterns and relationships it creates clusters and divides the dataset into those clusters. Now if a new data is fed to the model, it adds it to one of the created clusters. In unsupervised learning, as you might guess, the training data is unlabelled. The system tries to learn without a teacher.

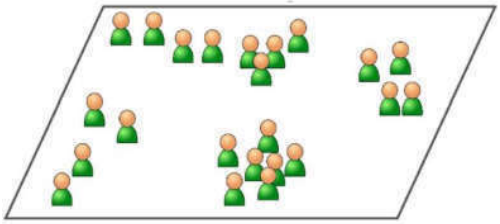


Fig. 9. Unsupervised Learning (training set)

Here are some of the most important unsupervised learning algorithms

- Clustering
- Visualization and dimensionality reduction
- Association rule learning.

For example, say you have a lot of data about your blog's visitors. You may want to run a clustering algorithm to try to detect groups of similar visitors. At no point do you tell the algorithm which group a visitor belongs to: it finds those connections without your help. For example, it might notice that 40% of your visitors are males who love comic books and generally read your blog in the evening, while 20% are young sci-fi lovers who visit during the weekends, and so on. If you use a hierarchical clustering algorithm, it may also subdivide each group into smaller groups. This may help you target your posts for each group.

[5] A group of similar things occurring closely together

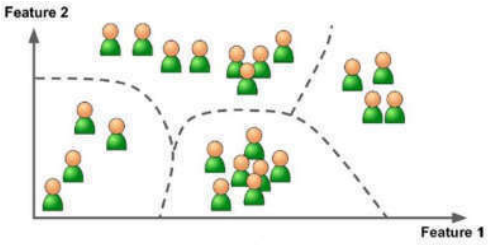


Fig. 10. Clustering in Unsupervised Learning

C. Reinforcement Learning

It is the ability of an agent to interact with the environment and find out what is the best outcome. It follows the concept of hit and trial method. The agent is rewarded or penalized with a point for a correct or a wrong answer, and on the basis of the positive reward points gained the model trains itself. And again, once trained it gets ready to predict the new data presented to it.[4]

VI. ROADMAP FOR BUILDING MACHINE LEARNING SYSTEMS

The diagram below shows a typical workflow diagram for using machine learning in predictive modelling.

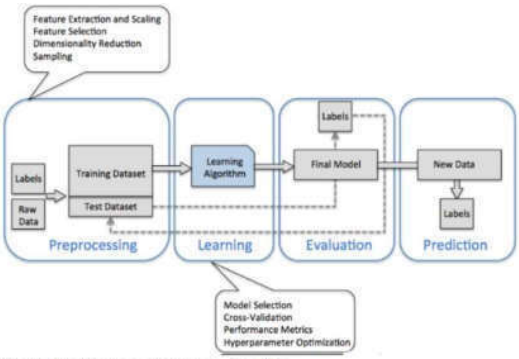


Fig. 11. Machine Learning System - Roadmap

A. Pre-processing - getting data into shape:

Raw data rarely comes in the form and shape that is necessary for the optimal performance of a learning algorithm. Thus, the pre- processing of the data is one of the most crucial steps in any machine learning.[5]

ISBN 978-81-939386-1-4

B. *Training and selecting a predictive model:*

It is essential to compare at least a handful of different algorithms in order to train and select the best performing model. But before we can compare different models, we first have to decide upon a metric[6] to measure performance. One commonly used metric is classification accuracy, which is defined as the proportion of correctly classified instances.[5]

C. *Evaluating models and predicting unseen data instances:*

After we have selected a model that has been fitted on the training dataset, we can use the test dataset to estimate how well it performs on this unseen data to estimate the generalization error. If we are satisfied with its performance, we can now use this model to predict new, future data. It is important to note that the parameters for the previously mentioned procedures such as feature scaling and dimensionality reduction are solely obtained from the training dataset, and the same parameters are later re-applied to transform the test dataset, as well as any new data samples the performance measured on the test data may be overoptimistic otherwise.[5]

VII. USING PYTHON FOR MACHINE LEARNING

Python is one of the most popular programming languages for data science and therefore, enjoys many useful add-on libraries developed by its great community.

Although the performance of interpreted languages, such as Python, for computation-intensive tasks is inferior to lower-level programming languages, extension libraries such as NumPy[7] and SciPy[8] have been developed that build upon lower layer Fortran and C implementations for fast and vectorized operations on multidimensional arrays. For machine learning programming tasks, we will mostly refer to the scikit-learn library, which is one of the most popular and accessible open source machine learning libraries as of today. [2]

Installing Python packages

Python is available for all three major operating systems—Microsoft Windows, Mac OS X, and Linux—and the installer, as well as the documentation, can be downloaded from the official Python website: *https://www.python.org*. and it is recommended to use the most recent version of Python 3 that is currently available, although most of the code examples may also be compatible with the version of Python >= 2.7.10.

If you decide to use Python 2.7 to execute the code examples, please make sure that you know about the major differences between the two Python versions. The additional packages that can be installed via the pip installer program, which has been part of the Python standard library since Python 3.3. After we have successfully installed Python, we can execute pip from the command line terminal to install additional Python packages:[2]

pip install SomePackage

Already installed packages can be updated via the --upgrade flag:

pip install SomePackage –upgrade

Another important library is NumPy. we will mainly use NumPy's multi-dimensional arrays to store and manipulate data. Occasionally, we will make use of pandas, which is a library built on top of NumPy that provides additional higher-level data manipulation tools that make working with tabular data even more convenient. To augment our learning experience and visualize quantitative data, which is often extremely useful to intuitively make sense of it, we will use the very customizable matplotlib library.

The example of version numbers of the major Python packages are listed below. Please make sure that the version numbers of your installed packages are equal to, or greater than, those version numbers to ensure the code examples run correctly:

- NumPy 1.9.1
- SciPy 0.14.0
- scikit-learn 0.15.2
- matplotlib 1.4.0
- pandas 0.15.2

[6] Measurement of a characteristic

[7] Package for scientific calculation in Python

[8] Another library for scientific computing in Python

VIII. CONCLUSION

In this paper, we have explored machine learning on a very high level and familiarized ourselves with the big picture and major concepts. We learned about What is Machine Learning, How and where we can use Machine learning, Different types of Machine Learning and that supervised learning is composed of two important subfields: classification and regression. While classification models allow us to categorize objects into known classes, we can use regression analysis to predict the continuous outcomes of target variables. Unsupervised learning not only offers useful techniques for discovering structures in unlabelled data, but it can also be useful for data compression in feature pre-processing steps.

We briefly went over the typical roadmap for applying machine learning to problem tasks, which can be used as a foundation for developing Machine learning systems. Eventually, we also saw how to set up our Python environment and install and updated the required packages to get ready to see machine-learning in action.

ACKNOWLEDGEMENT

We would like to acknowledge the contribution of all the people who have helped in reviewing this paper. We would like to give sincere thanks to Mrs. Subhashini Peneti, for her guidance and support throughout this paper. We would also like to thank our families and friends who supported us during writing this paper.

REFERENCES

[1] Hands-On Machine Learning with Scikit-Learn and TensorFlow, by - Aurélien Géron

[2] Python Machine Learning, by - Sebastian Raschka

[3] Machine Learning in Action, by Peter Harrington

[4] Machine Learning For Dummies, by - John Paul Mueller and Luca Massaron

[5] https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning#approaches

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Timing Enabled Proxy Re-Encryption Technique for E-Health Care Clouds in Keyword Search by Delegate

R.Sai Hanuma Chandrika,
M.Tech., Scholar,
Dept. of Computer Science & Engineering,
Malla Reddy College of Engineering,
Hyderabad.
rshchandrika@gmail.com

Dr. Gladson Maria Britto,
Professor,
Dept. of Computer Science & Engineering,
Malla Reddy College of Engineering,
Hyderabad.
gmbritto@gmail.com

Abstract— **An electronic health (e-health) record system is a novel application that will bring great convenience in healthcare. The privacy and security of the sensitive personal information are the major concerns of the users, which could hinder further development and widely adoption of the systems. The searchable encryption (SE) scheme is a technology to incorporate security protection and favorable operability functions together, which can play an important role in the e-health record system. In this paper, we introduce a novel cryptographic primitive named as conjunctive keyword search with designated tester and timing enabled proxy reencryption function (Re-dtPECK), which is a kind of a time-dependent SE scheme. It could enable patients to delegate partial access rights to others to operate search functions over their records in a limited time period. The length of the time period for the delegatee to search and decrypt the delegator's encrypted documents can be controlled. Moreover, the delegatee could be automatically deprived of the access and search authority after a specified period of effective time. It can also support the conjunctive keywords search and resist the keyword guessing attacks. By the solution, only the designated tester is able to test the existence of certain keywords. We formulate a system model and a security model for the proposed Re-dtPECK scheme to show that it is an efficient scheme proved secure in the standard model. The comparison and extensive simulations demonstrate that it has a low computation and storage overhead.**

I. Introduction

What is Secure Computing?

Computer security (Also known as cyber security or IT Security) is information security as applied to computers and networks. The field covers all the processes and mechanisms by which computer-based equipment, information and services are protected from unintended or unauthorized access, change or destruction. Computer security also includes protection from unplanned events and natural disasters. Otherwise, in the computer industry, the term security -- or the phrase computer security -- refers to techniques for ensuring that data stored in a computer cannot be read or compromised by any individuals without authorization. Most computer security measures involve data encryption and passwords. Data encryption is the translation of data into a form that is unintelligible without a deciphering mechanism. A password is a secret word or phrase that gives a user access to a particular program or system.



Diagram clearly explain the about the secure computing

Working conditions and basic needs in the secure computing:

If you don't take basic steps to protect your work computer, you put it and all the information on it at risk. You can potentially compromise the operation of other computers on your organization's network, or even the functioning of the network as a whole.

1. Physical security:

Technical measures like login passwords, anti-virus are essential. (More about those below) However, a secure physical space is the first and more important line of defense.

Is the place you keep your workplace computer secure enough to prevent theft or access to it while you are away? While the Security Department provides coverage across the Medical center, it only takes seconds to steal a computer, particularly a portable device like a laptop or a PDA. A computer should be secured like any other valuable possession when you are not present.

Human threats are not the only concern. Computers can be compromised by environmental mishaps (e.g., water, coffee) or physical trauma. Make sure the physical location of your computer takes account of those risks as well.

2. Access passwords:

The University's networks and shared information systems are protected in part by login credentials (user-IDs and passwords). Access passwords are also an essential protection for personal computers in most circumstances. Offices are usually open and shared spaces, so physical access to computers cannot be completely controlled.

To protect your computer, you should consider setting passwords for particularly sensitive applications resident on the computer (e.g., data analysis software), if the software provides that capability.

3. Prying eye protection:

Because we deal with all facets of clinical, research, educational and administrative data here on the medical campus, it is important to do everything possible to minimize exposure of data to unauthorized individuals.

4. Anti-virus software:

Up-to-date, properly configured anti-virus software is essential. While we have server-side anti-virus software on our network computers, you still need it on the client side (your computer).

5. Firewalls:

Anti-virus products inspect files on your computer and in email. Firewall software and hardware monitor communications between your computer and the outside world. That is essential for any networked computer.

6. Software updates:

It is critical to keep software up to date, especially the operating system, anti-virus and anti-spyware, email and browser software. The newest versions will contain fixes for discovered vulnerabilities.

Almost all anti-virus have automatic update features (including SAV). Keeping the "signatures" (digital patterns) of malicious software detectors up-to-date is essential for these products to be effective.

7. Keep secure backups:

Even if you take all these security steps, bad things can still happen. Be prepared for the worst by making backup copies of critical data, and keeping those backup copies in a separate, secure location. For example, use supplemental hard drives, CDs/DVDs, or flash drives to store critical, hard-to-replace data.

8. Report problems:

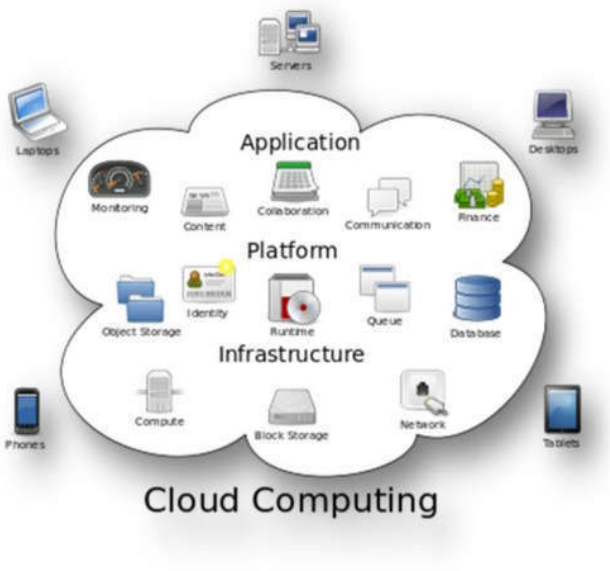
If you believe that your computer or any data on it has been compromised, your should make a information security incident report. That is required by University policy for all data on our systems, and legally required for health, education, financial and any other kind of record containing identifiable personal information.

Benefits of secure computing:

- **Protect yourself - Civil liability**: You may be held legally liable to compensate a third party should they experience financial damage or distress as a result of their personal data being stolen from you or leaked by you.
- **Protect your credibility - Compliance**: You may require compliancy with the Data Protection Act, the FSA, SOX or other regulatory standards. Each of these bodies stipulates that certain measures be taken to protect the data on your network.
- **Protect your reputation – Spam:** A common use for infected systems is to join them to a botnet (a collection of infected machines which takes orders from a command server) and use them to send out spam. This spam can be traced back to you, your server could be blacklisted and you could be unable to send email.
- **Protect your income - Competitive advantage:** There are a number of "hackers-for-hire" advertising their services on the internet selling their skills in breaking into company's servers to steal client databases, proprietary software, merger and acquisition information, personnel details*et al*.
- **Protect your business – Blackmail**: A seldom-reported source of income for "hackers" is to·break into your server, change all your passwords and lock you out of it. The password is then sold back to you. Note: the "hackers" may implant a backdoor program on your server so that they can repeat the exercise at will.
- **Protect your investment - Free storage:** Your server's harddrive space is used (or sold on) to house the hacker's video clips, music collections, pirated software or worse. Your server or computer then becomes continuously slow and your internet connection speeds deteriorate due to the number of people connecting to your server in order to download the offered wares.

What is cloud computing?

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the common use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation. Cloud computing consists of hardware and software resources made available on the Internet as managed third-party services. These services typically provide access to advanced software applications and high-end networks of server computers.



Structure of cloud computing

How Cloud Computing Works?

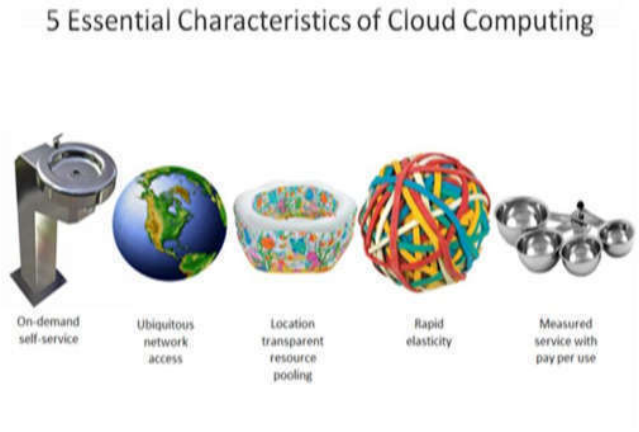
The goal of cloud computing is to apply traditional supercomputing, or high-performance computing power, normally used by military and research facilities, to perform tens of trillions of computations per second, in consumer-oriented applications such as financial portfolios, to deliver personalized information, to provide data storage or to power large, immersive computer games.

The cloud computing uses networks of large groups of servers typically running low-cost consumer PC technology with specialized connections to spread data-processing chores across them. This shared IT infrastructure contains large pools of systems that are linked together. Often, virtualization techniques are used to maximize the power of cloud computing.

Characteristics and Services Models:

The salient characteristics of cloud computing based on the definitions provided by the National Institute of Standards and Terminology (NIST) are outlined below:

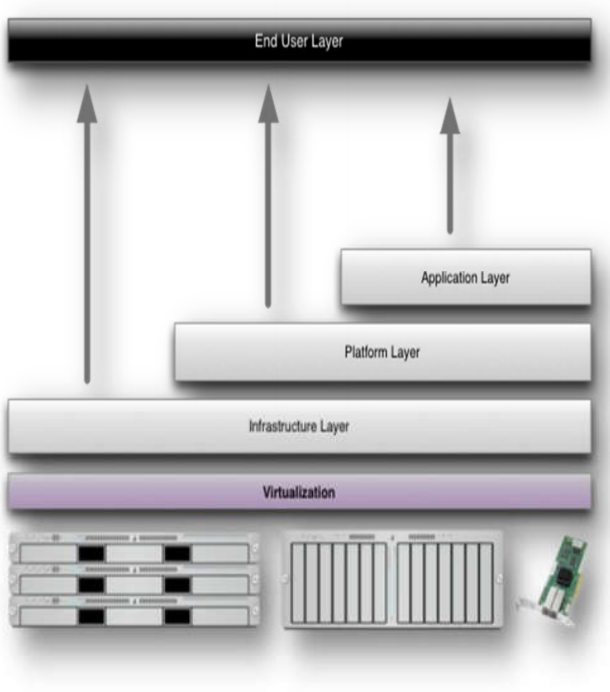
- **On-demand self-service**: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.
- **Broad network access**: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).
- **Resource pooling**: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location-independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or data center). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.
- **Rapid elasticity**: Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.
- **Measured service**: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be managed, controlled, and reported providing transparency for both the provider and consumer of the utilized service.



Characteristics of cloud computing

Services Models:

Cloud Computing comprises three different service models, namely Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). The three service models or layer are completed by an end user layer that encapsulates the end user perspective on cloud services. The model is shown in figure below. If a cloud user accesses services on the infrastructure layer, for instance, she can run her own applications on the resources of a cloud infrastructure and remain responsible for the support, maintenance, and security of these applications herself. If she accesses a service on the application layer, these tasks are normally taken care of by the cloud service provider.



Structure of service models

Benefits of cloud computing:

1. **Achieve economies of scale** – increase volume output or productivity with fewer people. Your cost per unit, project or product plummets.
2. **Reduce spending on technology infrastructure.** Maintain easy access to your information with minimal upfront spending. Pay as you go (weekly, quarterly or yearly), based on demand.
3. **Globalize your workforce on the cheap.** People worldwide can access the cloud, provided they have an Internet connection.
4. **Streamline processes.** Get more work done in less time with less people.
5. **Reduce capital costs.** There's no need to spend big money on hardware, software or licensing fees.
6. **Improve accessibility.** You have access anytime, anywhere, making your life so much easier!
7. **Monitor projects more effectively.** Stay within budget and ahead of completion cycle times.
8. **Less personnel training is needed.** It takes fewer people to do more work on a cloud, with a minimal learning curve on hardware and software issues.
9. **Minimize licensing new software.** Stretch and grow without the need to buy expensive software licenses or programs.
10. **Improve flexibility.** You can change direction without serious "people" or "financial" issues at stake.

Advantages:

1. **Price:** Pay for only the resources used.
2. **Security**: Cloud instances are isolated in the network from other instances for improved security.
3. **Performance:** Instances can be added instantly for improved performance. Clients have access to the total resources of the Cloud's core hardware.
4. **Scalability:** Auto-deploy cloud instances when needed.
5. **Uptime:** Uses multiple servers for maximum redundancies. In case of server failure, instances can be automatically created on another server.
6. **Control:** Able to login from any location. Server snapshot and a software library lets you deploy custom instances.
7. **Traffic:** Deals with spike in traffic with quick deployment of additional instances to handle the load.

II. Related Work

A. Conjunctive Keyword Search

Various constructions of public key encryption with conjunctive keyword search (PECK) over encrypted data have been proposed [8]–[10]. It allows the users to query multiple keywords at the same time [11], [12]. However, some of them such as the solution in [9] and [10] have high communication or computation cost. On the other hand, some schemes such as the solutions in [8] and [12] require an index list of the queried keywords when a trapdoor is generated, which will leak information and impair the query privacy.

B. Searchable Encryption With Designated Tester

In practice, the size of a keyword space is always no more than its polynomial level. An attacker is possibly to launch dictionary attacks or off-line keyword guessing attacks (KG attacks) to exploit the hidden keywords. The EHR keywords

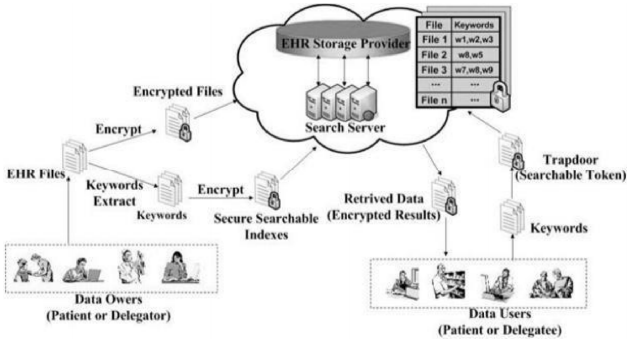


Fig. 1. System Model.

are usually selected from a small space, especially the medical terminology. If an adversary finds that the trapdoors or encrypted indexes have lower entropies, the KG attacks could be launched if the adversary endeavors to guess the possible candidate keywords. Byun *et al.* [19] and Yau *et al.* [20] have broken several classical schemes by the KG attacks.

In order to resist the threats, the concept of PEKS with designated tester (dPEKS) is proposed in [21]–[25]. Only a designated tester, which is usually the server, is capable to carry on the test algorithm. The enhanced security models [7], [26], [27] have also been put forward. However, they could not support multiple keywords query or delegate search function.

C. Proxy Re-Encryption With Public Keyword Search

Proxy re-encryption (PRE) enables a proxy with a re-encryption key to convert a ciphertext encrypted by a delegator's public key into those that can be decrypted by delegatee's private key. Proxy re-encryption with public keyword search (Re-PEKS) [13]–[15] has introduced the notion of keyword search into PRE. The users with a keyword trapdoor can search the ciphertext while the hidden keywords are unknown to the proxy. The limitation on the schemes in [13]–[15] is that only one keyword will be allowed to search in the encrypted documents. Later, Wang *et al.* [16] has suggested an improved scheme to support the conjunctive keyword search function. All these Re-PEKS schemes in [13]–[16] are proved secure in random oracle model. Nevertheless, it is shown in [17] and [18] that a proof in random oracle model may probably bring about insecure schemes.

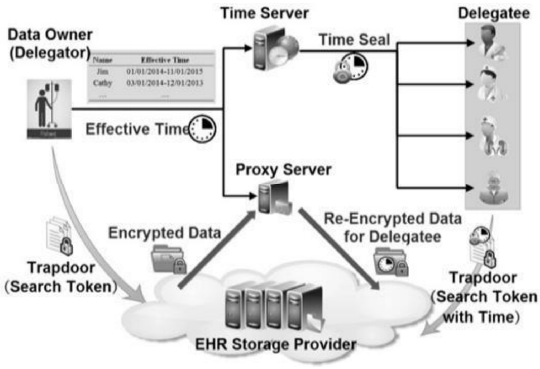


Fig. 2. Timing Enabled Proxy Re-encryption Searchable Encryption Model.

three types of entities: an information owner, users and a data center. The data owner wants to store his private EHR files on a third-party database. He extracts keywords from the EHR files and encrypts those plaintext keywords into the secure searchable indices. The EHR files are encrypted to ciphertext. Then, those information are outsourced to the data center.

A data center consists of an EHR storage provider and a search server. The storage provider is responsible for storing data and search server performs search/add/delete operations according to users' requests. A user generates a trapdoor to search the EHR files using his private key and sends it to the search servers. After receiving the request, the search servers interact with the EHR storage provider to find the matched files and returns those retrieved information to the user in an encrypted form.

In Fig. 2, the timing enabled proxy re-encryption searchable encryption model is shown. In this model, we highlight the implementation of the time controlled function. The data owner acting as a delegator sends a list of delegation effective time periods for his delegatees to the time server and the proxy server. The entry of the list contains the identity of each delegatee and the effective time period, such as "Jim, 01/01/2014 – 11/01/2015". It indicates that the delegatee Jim is authorized to issue queries and perform decryption operations on the encrypted data of the data owner from Jan. 1th, 2014 to Nov. 1th, 2015. After receiving the list, the time server generates a time seal for each delegatee, which is

The time controlled PRE has been addressed in [28]–[30]. It desires to encrypt a message for multiple recipients with the same release time. However, the schemes in [28] and [30] foist the data owner to determine the release time at the beginning of encryption algorithm. Only one release time is set for all recipients rather than disparate time for different users, which could not fulfill the need for uniqueness. Another shortcoming is that it needs a large computation cost in both encryption and re-encryption phases [29].

III. Problem Formulation

A. System Model

In Fig. 1, the environment of the proposed Re-dtPECK scheme for the EHR cloud system is presented. There are transmitted to individuals. The time seal is a trapdoor of an effective time period and concealed by the private key of the time server. In the re-encryption operation, the proxy server will encapsu-late the effective time into the re-encrypted ciphertext. In order to reduce computing cost, the proxy server will not re-encrypt the ciphertext until they are accessed, which is so called lazy re-encryption mechanism [31]. In the query phase, the data owner can conduct ordinary search operations with his own private key. However, the delegatee has to generate a keywords trapdoor with the help of the time seal. The cloud data server will not return the matched files unless the effective time encapsulated in the time seal accords with the time in the re-encrypted ciphertext, which is different from traditional proxy re-encryption SE schemes.

B. Threat Model

The EHR data server is deemed as semi-trusted, who is honest to search information for the benefits of users but curious to spy out the private information of the patients. On the other hand, malicious outside attacker could eavesdrop and analyze the information transferred in public channel, such as the encrypted indexes and trapdoors. He intends to infer privacy information according to these data. Furthermore, the revoked delegatees may try to access data beyond the designated time period using their private keys. As most of the storage and search work are completed by the data server, it is assumed that the data server will not collude with the malicious outside attacker or revoked delegatees.

C. Design Goals

Our Re-dtPECK scheme for EHR cloud is designed to achieve the following goals.

- *1)* ***Authority delegation***. The proposed SE scheme should allow data-owner-enforced authority delegation, i.e. the data owner could delegate his search right to other users without revealing his private key.
- *2)* ***Time controlled revocation***. An important design goal is to enable time controlled access right revocation. The delegation appointment will terminate when the preset effective time period disagrees with the current time. It should prevent the authorized user from accessing the records overtime.
- *3)* ***Diverse delegation times for different users***. Another challenge of the system is to achieve owner-defined

disparate access time periods for different delegatees. The data owner himself will not be constrained by the time.

- *4)* ***Security goals***. The privacy concerns of this secure search system are summarized as follows. 1) *keyword semantic security*: as a Re-dtPECK scheme is proposed, we will prove it indistinguishable against chosen keywords chosen time attack (IND-CKCTA). 2) *resist KG attacks*: since the EHR keywords are always chosen from a small space, the related searchable encryption schemes maybe vulnerable to offline KG attacks. The proposed scheme should resist such attack. 3) *standard model*: it is well known that security proved in standard model is stronger than that in random oracle model. This security property guarantees a higher security level.

IV. The Proposed Re-dtPECK

TABLE I
Summary of Notations

Notation	Description
θ	Delegation indicator
GP	Global parameter
sk_S, pk_S	Private and public keys of data server
sk_{TS}, pk_{TS}	Private and public keys of time server
sk_R, pk_R	Private and public keys of user
R_i, R_j	Delegator and Delegatee
W, Q	Keyword set
S_T	Time seal: trapdoor of effective time
$C_I (C_J)$	Ciphertext for delegator (delegatee)
$T_{Q,I} (T_{Q,J})$	Trapdoor of delegator (delegatee) on Q

- *GlobalSetup(k)*: Taking a security parameter k as an input, this function generates a global parameter GP.
- *KeyGen$_{Ser}$ (GP)*: Taking GP as an input, this algorithm generates a private and public key pair (sk_S, pk_S) for the data server.
- *KeyGen$_{Rec}$ (GP)*: Taking a global parameter GP as an input, this function generates a private and public key pair (sk_R, pk_R) for the receiver.
- *KeyGen$_{TS}$ (GP)*: Taking a global parameter GP as an input, this function generates a private and public key pair (sk_{TS}, pk_{TS}) for the time server.
- *dPECK($GP, pk_S, pk_{R_i}, sk_{R_i}, W$)*:Taking GP, pk_S, pk_{R_i}, sk_{R_i} and a keyword set $W = (w_1, \cdots, w_l)$ as the inputs, the function returns a ciphertext C_I of W for R_i.
- *Trapdoor(GP, pk_S, sk_{R_i}, Q)*: Taking GP, pk_S, sk_{R_i} and a keyword query for $Q = (w_1, \cdots, w_m)$, $m \leq l$as the inputs, it outputs a trapdoor $T_{Q,I}$ for Q generated by R_i.
- *Test(GP, $T_{Q,I}$, sk_S, C_I)*: Taking GP, $T_{Q,I}$, sk_S and a ciphertext C_I of W as the inputs, the function returns '1' if W includes Q and '0' otherwise.

If the delegation indicator θ equals to 1, the following operations are executed.

- *ReKeyGen(GP, sk_{R_i}, sk_{R_j})*: Taking GP, sk_{R_i}, sk_{R_j} as the inputs, the algorithm outputs re-encryption key $rk_{R_i \to R_j}$.
- *Re − dtPECK(GP, $rk_{R_i \to R_j}$, C_I, pk_i, pk_j, pk_{TS}, T)*: Taking GP, $rk_{R_i \to R_j}$, C_I, pk_i, pk_j, pk_{TS}, T as the inputs, the algorithm outputs a re-encryption ciphertext C_J.
- *TimeSeal(GP, sk_{TS}, T, pk_{R_i}, pk_{R_j})*: Taking GP, sk_{TS}, T, pk_{R_i}, pk_{R_j} as the inputs, it outputs a time seal S_T for user R_j in order to search user R_i's encrypted data.
- *Trapdoor$_R$ (GP, pk_S, sk_{R_j}, Q, S_T)*: Taking GP, pk_S, sk_{R_j}, Q, S_T as the inputs, it outputs a trapdoor $I_{Q,J}$ for R_j.
- *Test$_R$ (GP, $T_{Q,J}$, sk_S, C_J)*: Taking GP, $T_{Q,J}$, sk_S and C_J as inputs, the function returns '1' if W includes Q and the effective time contained in $T_{Q,J}$ is accordance with the time encapsulated in C_J. Otherwise, it outputs '0'.

In this subsection, we formally define the conjunctive keyword search with a designated tester and the timing enabled proxy re-encryption function (Re-dtPECK). Then, we describe a concrete Re-dtPECK scheme with a detailed workflow and derive the correctness of the scheme. The notations presented in this paper are summarized in Table I.

A. Definition of Re-dtPECK

The Re-dtPECK scheme consists of following algorithms with an indicator θ. When its value is 1, the delegation function will be activated. Otherwise, the proxy re-encryption will not be enabled.

Correctness: For any security parameter k, we have

$$Test \begin{cases} 1, & W \subseteq Q \\ \to & \\ 0, & otherwise \end{cases}, \quad Test_R \begin{cases} 1, & W \subseteq Q, T_1 = T_2 \\ \to & \\ 0, & otherwise \end{cases}$$

where T_1 is the effective time contained in $T_{Q,J}$ and T_2 is the effective time encapsulated in C_J.

The semantic security model is defined in Appendix. A.

B. Construction of Re-dtPECK

In the system, the EHR documents of the patients are encrypted by a symmetric encryption algorithm and the symmetric key is encapsulated with the patient's public key pk_A by the key encapsulation mechanism. The algorithms in the following focus on the searchable keywords encryption and the timing controlled delegation function. The time T chosen from the time space $T \in \{0, 1\}$ has the form "Month/Day/Year-Month/Day/Year" to set the starting time and closing time of the delegation. For example, "05/01/ 2013-12/01/2014" means that the delegatee is authorized to operate on delegator's secret data from May 1st, 2013 to Dec. 1st, 2014. In this proposal, a hash function H is defined as $H : \{0, 1\}^* \to Z^*_p$.

- *GlobalSetup(k)*: Let k be the security parameter. Choose a random generator g of group G. The global parameter is $GP = g$.
- *KeyGen$_{Ser}$ (GP)* This algorithm randomly chooses $V \in G$ and $x \in Z^*_p$, computes $X = g^x$ and outputs the data server's key pair $pk_S = (V, X)$, $sk_S = x$.
- *KeyGen$_{Rec}$ (GP)*: For user R_i, choose a random value $y_i \in Z^*$ and compute $Y_i = g^{y_i}$. The public and private key pair of R_i is $pk_{R_i} = Y_i$ and $sk_{R_i} = y_i$.

- *KeyGen$_{TS}$ (GP)*: This algorithm selects $\tau \in Z^*_p$ randomly and outputs a public and secret key pair for the time server as $pk_{TS} = g^{\tau}$, $sk_{TS} = \tau$.
- *dPECK(GP, pk_S, pk_{Ri}, sk_{Ri}, W)*: The sender selects a keyword set $W = (w_1, \cdots, w_l)$ from the file to be outsourced. An l-degree polynomial $Y(x) = \eta_l x^l + \eta_{l-1} x^{l-1} + \cdots + \eta_1 x + \eta_0$ should be constructed to make $y_i H(w_1), \cdots, y_i H(w_l)$ to be l roots of the equation $Y(x) = 1$. Then data sender randomly chooses $s, r \in Z^*_p$ and computes $C_1 = t \cdot e(g, g)^{-r}$, $C_2 = g^s$, $B_\psi = g^{r \cdot \eta_\psi}$, for $0 \le \psi \le l$, where $t = e(X, V)^s$. It then outputs $C_I = (C_1, C_2, B_0, B_1, \cdots, B_l)$. It is clear that the bilinear pairing operations $e(g, g)$ and $e(X, V)$ can be precomputed so that no additional time will be consumed.

The delegator R_i sends out a delegation notice to the trusted third party, time server, proxy server, data server and delegatee R_j. The delegation information is R_i, R_j, T, $Sig_{skRi}(R_i, R_j, T)$, in which T denotes the designated delegation time period and $Sig_{skRi}(R_i, R_j, T)$ is a signature of the information (R_i, R_j, T) generated by delegator R_i used to verify the authenticity. *Sig* is a cryptographic secure signature algorithm without a specification in this scheme. The signature can be verified by using the public key of R_i. The delegation request could be rejected if the signature is forged. Otherwise, the following algorithms will be executed.

- *ReKeyGen(GP, sk_{Ri}, sk_{Rj})*: Taking GP, $sk_{Ri} = y_i$, $sk_{Rj} = y_j$ as the inputs, this algorithm outputs re-encryption key $rk_{Ri \to Rj} = y_i / y_j$.
- *ReEncdtPECK(GP, $rk_{Ri \to Rj}$, C_I, pk_i, pk_j, pk_{TS}, T)*: This algorithm computes $C_3 = (g^{-H(T, pk_i)}_{\psi} pk_{TS})^{\mu}$, $C_4 = Y_j^{\mu}$, $C_5 = g^{\mu}$, $B_\psi = B_{\psi\, Ri \to Rj}$, for $0 \le \psi \le l$ and outputs a Re-dtPECK ciphertext $C_J = (C_1, C_2, C_3, C_4, C_5, B_0, B_1, \cdots, B_l)$ for delegatee R_j.
- *TimeSeal(GP, $skTS$, T, pkR_i, pkR_j)*: Taking the time server's private key $skTS = \tau$, the public key pkR_i, pkR_j and a designated time period $T \in \mathcal{T}$ as the inputs, it returns a time seal $S_T = (r_T, h_T)$, where r_T is a random element from Z^* and $h_T = (Y_j \; g^{-r_T})^{1/(\tau - H(T, pk_i))}$.
- *TrapdoorR (GP, pk_S, skR_j, Q, S_T)*: It computes $T_{Q,1} = (T_{Q,1}, T_{Q,2}) = (g^{r_T}, (h_T)^{y_j})$, $T_{Q,2} = g^{\zeta}$, $T_{Q,\psi} = g^{m_{\psi}} \cdot T_{Q,-1}^{(y_i)} \cdot \prod_{\mu=1}^{H(w_{\gamma\mu})} \cdot X^{\zeta}$ where $0 \le \psi \le l$, $Q_w = w(w_{\gamma 1}, \cdots, w_{\gamma m})$, $m \le l$. Then, the algorithm outputs the trapdoor $T_{Q,I} = (T_{Q,-1}, T_{Q,-2}, T_{Q,\psi})$ for $0 \le \psi \le l$.
- *Test(GP, $T_{Q,I}$, skS, C_I)*: After receiving the trapdoor $T_{Q,I}$ and the ciphertext C_I, the server computes $t = e(C_2, V)^{x}$ and checks whether the following equation holds $C_1^{T_{Q,-1}} \prod_{\psi=0}^{l} e(B_\psi, T_{Q,\psi}/(T_{Q,-2})^{x}) = t^{T_{Q,-1}}$. If the equation holds, it outputs 1. Otherwise, it outputs 0. Appendix.C demonstrates the correctness of the scheme.

The designs goals in section III.C can be fulfilled based on the following aspects. 1) The authority delegation is realized mainly by proxy re-encryption mechanism. The proxy server makes use of the re-encryption key to transform the ciphertext encrypted by delegator's public key into another form, which can be searched by the delegatee using his own private key.

If delegation indicator θ equals to 1, execute the following operations.

2) The delegatee will be divested of the search authority when the effective time expires. In order to achieve the time controlled access right revocation, the predefined time information is embedded in the re-encrypted ciphertext with a time seal. With the help of the time seal, the delegatee is able to generate a valid delegation trapdoor by *TrapdoorR* algorithm. If the time information hidden in the re-encrypted ciphertext is inconsistent with that in the delegation trapdoor, the equation in *TestR* algorithm will not hold. Moreover,

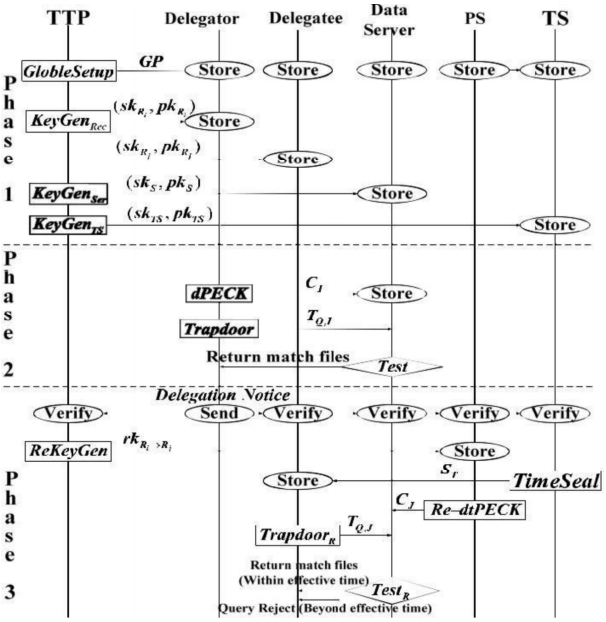


Fig. 3. Workflow of Re-dtPECK.

3) Since the time server is able to generate different time seals according to the requirement of the data owner, different delegation time periods can be set for different users through distributing the distinct time seals. The patient himself will not be constrained by the effective time period because the limitation is made in the delegation phase rather than the original encryption phase. 4) The security of this scheme is analyzed in Section V.

C. Workflow of Re-dtPECK

The detailed workflow is shown in Fig. 3 to illustrate the working procedure of the Re-dtPECK system in the EHR cloud. There are six entities to participate in the interactive process

including a trusted third party (TTP). For instance, the Veterans Health Administration (VHA) is assumed to work as a TTP, who is trusted by clinics, hospitals, patients and doctors. A delegator is supposed to be Joe, who is a chronic heart failure patient. The EHR files of Joe are stored on a data server in the cloud in a protected form. Joe went to Hospital A for the cardiac treatment since Feb. 1st, 2014. He desires to designate the cardiologist Dr. Donne from Hospital A to be his delegatee for convenient EHR data access. Since Joe plans to transfer to Hospital B after June 1st and he hopes that Dr. Donne is not able to inquiry his EHR after that time. Then, Dr. Donne is granted a time-constrained which are disseminated to delegator Joe, delegatee Dr. Donne, the EHR cloud server, the PS and the TS. The TTP also generates pairs of private and public key for Joe, Dr. Donne, the cloud server and the TS by running $KeyGen_{Rec}$, $KeyGen_{Ser}$, $KeyGen_{TS}$ algorithms.

In phase 2, EHR files are produced during Joe's therapeutic process. The encrypted EHR indices and documents will be generated using the dPECK algorithm and stored at the cloud data server. As the medical data accumulated, Joe may intend to search on his encrypted EHR files. He uses a keyword set Q to describe the health record file that he wants to find. Then, he runs *Trapdoor* algorithm to generate a trapdoor for keyword set Q and sends the trapdoor to cloud server. After receiving the search query, the cloud server runs test algorithm with the cloud server's private key and returns all the files that contain Q.

If the delegation indicator θ equals to 1, phase 3 will be executed. Joe sends a delegation notice to the TTP, PS, TS, delegatee and data server together with a signature signed by Joe. The effective delegation time of PHI access delegation for delegatee is specified. It means that the patient Joe has delegated the access rights to Dr. Donne. The recipients will verify the signature using the public key of Joe. In this system, the signature algorithm will not be specified. But there is a requirement on the algorithm that the signature scheme should be strongly unforgeable. The notice will be rejected if the signature fails the verification. If it is verified true, the TTP runs ReKeyGen algorithm to generate a re-encryption key and send it to the PS secretly. The TS runs *TimeSeal* algorithm to generate a time seal for delegatee. When Joe's PHI data is accessed by the Dr. Donne, the PS will run Re-dtPECK algorithm to encapsulate the effective time period into re-encrypted ciphertext. If the current time is not in accordance with the effective time period, the PS will not do the re-encryption operation for Dr. Donne. Suppose Dr. Donne is responsible to perform a heart bypass operation for Joe on Mar. 12nd, 2014. It is necessary for Dr. Donne to gain access to Joe's historical health record in order to prepare for the operation. Dr. Donne runs delegation trapdoor generation algorithm to get a valid trapdoor, which is used to conduct search query on Joe's PHI data. After receiving the query, cloud server runs the delegation test algorithm. The search query will be rejected if the current time beyond the effective time period. If the delegation test algorithm holds and current time is within the effective time period, the matched files will be returned to Dr. Donne.

V. Security Analysis

A. Hardness Problem

authority to access the protected health information (PHI) of the patient Joe. The time server (TS) will generate a time seal for Dr. Donne to ensure that he is able to access to Joe's PHI during the period of Feb. 1st- May, 30st, 2014. The proxy server (PS) is responsible to encrypt Joe's PHI to a re-encrypted form so that Dr. Donne can search on those records with his own private key. The Re-dtPECK system can be divided into three phases.

In phase 1, the TTP initializes the system by executing *GlobalSetup* algorithm and generates the global parameters,

The following hardness problems are used in the security proof in next subsection.

Definition 1 (Truncated Decisional l -Augmented Bilinear Diffie-Hellman Exponent Problem (Truncated Decisional l-ABDHE)): Given a group G of prime order q, the algorithm randomly chooses generators g, g from G and $\alpha \in Z_q^*$ at random. Given a tuple $Tu = (g, g_1, g_2, g_{l+2}, g, g_1, \cdots, g_l)$ and $Z \in G_1$, where we use g_i and g_i to denote g^{α_i} and $(g)^{\alpha_i}$, the truncated decisional l-ABDHE problem is to decide whether Z equals to $e(g, g_{l+1})$ or to a random element of G_1.

Definition 2 (Decisional Bilinear Diffie-Hellman Problem (DBDH)): Given a group G of prime order q, the algorithm randomly chooses generators g from G and $a, b, c \in Z_q^*$ randomly. Given a tuple $Tu = (g^a, g^b, g^c)$ and $Z \in G_1$, the DBDH problem is to decide whether Z equals to $e(g, g)^{abc}$ or to a random element of G_1.

Definition 3 (Decisional Diffie-Hellman Problem (DDH)): Given a group G of prime order q, the algorithm randomly chooses generators g from G and $\alpha, \beta \in Z_q^*$ at random. Given a tuple $Tu = (g, g^\alpha, g^\beta)$ and $Z \in G$, the DDH problem is to decide whether Z equals to $g^{\alpha\beta}$ or to a random element of G.

B. Security

In this section, the proposed solution is proved IND-CKCTA and IND-KGA (indistinguishable against keyword guessing attack).

- ***Confidentiality***: The notion of confidentiality of the EHR in this paper means that the private documents of users must be kept secret from both unauthorized system visitors and the EHR cloud service provider. By this scheme, the health information is protected by means of a strong encryption primitive. The indexes of the conjunctive keywords are encrypted by the dPECK or Re-dtPECK algorithms before uploaded to the cloud server. The service provider could not recover the plaintext of the encrypted data. The keyword extraction from EHR is controlled by the patient and encrypted locally with patient R_i 's own secret key. The ciphertext can be derived with R_i 's private key skR_i . However, the server could not get any information about the patients' private keys for generating trapdoors and decrypting the protected documents.

Theorem 1: The proposed Re-dtPECK scheme is IND-CKCTA in the standard model assuming that the truncated decisional l-ABDHE problem and the DBDH problem are intractable.

Proof: The proof of Theorem 1 is in Appendix. B.

Theorem 2: The proposed Re-dtPECK scheme is IND-KGA secure in the standard model assuming that the DDH problem are intractable.

Proof: The proof of Theorem 2 is in Appendix. B.

Hint: There are two types of attackers in IND-CKCTA security proof including the server and the outside attacker. IND-CKCTA ensures that the server could not distinguish the ciphertext encrypts from keyword set and time if the trapdoor for given keywords and time is not obtained. On the other hand, the outside attacker could not make decisions about the ciphertext of certain keywords and time without the server's private key even though all the trapdoors for the other keywords and times are available.

IND-KGA guarantees that the attackers including the server attackers and outside attackers could not find the relationship between the given trapdoor and the challenge keywords even though other trapdoors for both delegator and delegatee can be obtained.

Since the first PEKS scheme proposed by Boneh *et al.* [4], almost all searchable encryption schemes in public key setting

The comparison in Table II indicates that the proposed scheme has versatile functions and higher security level.

TABLE II
FEATURES COMPARIS ON WITH OTHER SCHEMES

Scheme	F1	F2	F3	F4	F5	F6	F7	F8
[4]	No	No	No	No	No	No	Yes	Yes
[6]	No	No	Yes	No	--	No	Yes	Yes
[8]	No	Yes	Yes	No	Yes	No	Yes	Yes
[9]	No	Yes	No	No	No	No	Yes	Yes
[10]	No	Yes	No	No	No	No	Yes	Yes
[13]	No	No	No	Yes	No	No	Yes	Yes
[14]	No	No	No	Yes	No	No	Yes	Yes
[15]	No	No	No	Yes	No	No	Yes	Yes
[16]	No	Yes	No	Yes	No	No	Yes	Yes
[21]	No	No	Yes	No	No	No	Yes	Yes
[22]	No	No	Yes	No	Yes	No	Yes	Yes
[23]	No	No	Yes	No	No	No	Yes	Yes
[24]	No	No	Yes	No	--	No	Yes	Yes
[25]	No	No	Yes	No	Yes	No	Yes	Yes
[7]	No	No	Yes	No	Yes	No	Yes	Yes
[28]	Yes	No	No	No	Yes	No	Yes	Yes
[29]	Yes	No	No	No	--	No	Yes	Yes
[30]	Yes	No	No	No	No	No	Yes	Yes
[33]	No	Yes	--	No	--	Yes	No	No
[34]	No	Yes	--	No	--	Yes	No	Yes
Ours	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**No**	**Yes**	**Yes**

Note:
F1: Time control F2: Conjunctive keywords F3: Against KG attack
F4: Proxy search F5: Standard model F6: Boolean query
F7: No key sharing F8: Dynamic data change

- ***Time Control***: In privacy-preserving keyword query systems, the time controlled function will enable the EHR data owner to flexibly disseminate his access right to another user in a specified period of time. The delegation revocation can be achieved even when the EHR data owner is offline. Only some proxy re-encryption schemes in [28]–[30] without a query function have addressed it. Unfortunately, all the available searchable encryption schemes cannot hold this feature.
- ***Conjunctive Keywords***: Compared with the single keyword search, the conjunctive keyword search function provides the users more convenience to return the accu-rate results that fulfills users' multiple requirements. The users do not have to query an individual keyword and rely on an intersection calculation to obtain what they needs. To the best of our knowledge, there is no existing proxy re-encryption searchable encryption scheme could provide the conjunctive keywords search capability with-out requiring a random oracle. Our scheme has solved this open problem. The scheme in [16] could provide both the conjunctive keywords search and the delegation function. Unfortunately, it is proved in the random oracle (R.O.) model, which greatly impairs the security level.
- ***Against off-line KG attack***: Up to date, there is no effective approach to prevent the eavesdropping attacks over a public communication channel. An outside attacker could easily implement an off-line KG attack by monitoring the information channel between the patient and the data center if no powerful countermeasure is applied. Most of the existing schemes have not taken KG attacks into consideration. The schemes in [4], [9], [10], [13]–[16], and [28]–[30] could not resist KG attacks. The proposed scheme is immune to KG attacks, which is proved in Theorem 2.
- ***Proxy***: The proxy re-encryption technology is practical in EHR systems. It will greatly facilitate patient delegating the search and access rights. Schemes in [4], [6]–[10], [21]–[25], [28]–[30], [33] and [34] could not provide the proxy re-encryption searchable encryption function to the users.
- ***Standard model*:** There is no formal security proof on the schemes in [6], [24], and [29]. The schemes in [4], [9], [10], [13]–[16], [21], [23], and [30] have been proved to be secure relying on R.O. model. However, it could not demonstrate the security functionality in a real life. It is shown in [17] and [18] that the schemes might lead to insecure schemes when the random oracles are substituted with concrete hash functions. Thus, it is better to design a secure primitive that does not rely on the random oracle model. Our scheme is proved secure based on the standard model, which has a higher level security.
- ***No key sharing in multi-client setting***: The schemes in [33] and [34] are constructed in symmetric search-able

do not include test queries in security proof. The reason is that the test algorithm can be run if the keyword trapdoor and ciphertext are obtained. In fact, the adversary is able to get the trapdoors through trapdoor queries for the delegator or delegatee. The target ciphertext can be obtained in challenge phase. In PEKS schemes without designated tester, the test algorithm can be run by any attacker. In this work, the test algorithm can only be executed by the data server using his private key, which is the solid meaning of "designated tester".

VI. PERFORMANCE ANALYSIS

Security level, efficiency and the utility function are important indicators to evaluate whether a scheme is suitable for the privacy–preserving in the EHR cloud storage. The proposed Re-dtPECK will be compared with other relevant schemes according to these indicators. A simulation result on an experimental test-bed is also provided to measure the performance of Re-dtPECK scheme.

A. Comparison

In this subsection, we have compared our proposed Re-dtPECK scheme with other searchable encryption schemes in [4], [6]–[10], [13]–[16], [21]–[25], [33],
and [34] and the relevant proxy re-encryption schemes in [28]–[30] in terms of functionality, communication and computation overhead. The comparison results are shown in Table II and Table III.

encryption (SSE) setting, which allows a single user to implement upload or query operations. In order to be used in multi-client setting, the scheme has to make the symmetric key be shared by authorized users. Other schemes built in PEKS setting do not suffer such problem.

- ***Dynamic data change***: The scheme in [34] can support additions as well as deletions on encrypted data without using the same symmetric key, while the scheme in [33] cannot. The other schemes that are built in PEKS set-ting can also implement additions and deletions oper-ations through using data owner's public and private keys.

Scheme	Communication Overhead							Computation Overhead						
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
[8]	Yes	2	1	l+2	l+2	--	--	t_{Pair}+2t_{Exp}	(l+1)t_{Exp}	--	2t_{Exp}	--	3t_{Pair}	--
[9]	Yes	3l+3	3l+2	2l+2	2l+1	--	--	t_{Pair}+(3l+2)t_{Exp}	(3l+2)t_{Exp}	--	(5l+1)t_{Exp}	--	(2l+1)t_{Pair}	--
[10]	Yes	l+5	1	2l+4	l+3	--	--	t_{Exp}	t_{Pair}+(2l+4)t_{Exp}	--	(3l+5)t_{Exp}	--	(2l+1)t_{Pair}	--
[12]	Yes	1	1	l+1	l+1	--	--	2t_{Exp}	(2l+1)t_{Exp}	--	t_{Exp}	--	2t_{Pair}	--
[13]	No	1	1	7	1	7	--	t_{Exp}	2t_{Pair}+5t_{Exp}	t_{Exp}	t_{Exp}	--	t_{Pair}	--
[14]	No	1	1	2	2	2	--	t_{Exp}	t_{Pair}+3t_{Exp}	t_{Exp}	3t_{Exp}	--	t_{Pair}	--
[15]	No	5	5	7	4	4	--	5t_{Exp}	3t_{Pair}+3t_{Exp}	t_{Pair}+2t_{Exp}	2t_{Exp}	--	t_{Pair}	--
[16]	Yes	l+3	3	2l+4	3	l+3	--	2(l+3)t_{Exp}	(4l+3)t_{Exp}	2t_{Pair}	t_{Exp}	--	2t_{Pair}	--
[22]	No	6	1	4	2	--	--	t_{Exp}	t_{Pair}+6t_{Exp}	--	4t_{Exp}	--	2t_{Pair}	--
[23]	No	1	1	2	2	--	--	2t_{Exp}	t_{Pair}+2t_{Exp}	--	3t_{Exp}	--	t_{Pair}	--
[24]	No	3	1	2	2	--	--	4t_{Exp}	3t_{Pair}+2t_{Exp}	--	3t_{Exp}	--	t_{Pair}	--
[25]	No	1	1	2	2	--	--	4t_{Exp}	4t_{Pair}+5t_{Exp}	--	3t_{Exp}	--	t_{Pair}	--
[7]	No	l+3	l+3	7	2	--	--	(l+3)t_{Exp}	3t_{Pair}+8t_{Exp}	--	t_{Exp}	--	3t_{Pair}	--
[28]	No	1	1	7	--	9	--	t_{Exp}	4t_{Pair}+8t_{Exp}	4t_{Exp}	--	--	--	--
[29]	No	3	2	N+2	--	3N+2	--	5t_{Exp}	t_{Pair}+(N+1)t_{Exp}	t_{Pair}+(6N+1)t_{Exp}	--	--	--	--
[30]	No	1	1	7	--	7	--	t_{Exp}	t_{Pair}+7t_{Exp}	7t_{Exp}	--	--	--	--
Ours	Yes	1	1	l+3	l+3	l+6	l+4	t_{Exp}	(l+4)t_{Exp}	(l+5)t_{Exp}	(l+1)t_{Exp}	(l+2)t_{Exp}	(l+2)t_{Pair}	(l+4)t_{Pair}

Note: T1: Conjunctive keywords; T2:Public key size; T3: Private key size (user); T4: Ciphertext size (delegator); T5: Tradoor size (delegator); T6: Re-encryption ciphertext size; T7: Trapdoor size (delegatee); T8: Key Generation; T9: Encryption; T10: Re-encryption; T11: Trapdoor (delegator); T12: Trapdoor (delegatee); T13: Test (delegator); T14: Test (delegatee)
l: size of keyword set. N: number of conjunctive clauses in the access structure [29]
t_{Exp}: execution time for exponentiation operation; t_{Pair} : execution time for bilinear pairing operations.

Thus, the proposed scheme has various useful functions and has stronger security functionality than those of most of the existing searchable encryption schemes.

The communication overhead comparison in Table III shows that the proposed Re-dtPECK scheme has a small size of public key and private key.

- The sizes of the public keys in the schemes in [7], [9], [10], and [16] are linear with the number l of the conjunctive keywords, while the size of the public keys in this Re-dtPECK scheme is only one element in group G.
- The size of the private keys in the scheme in [7] and [9] is linear with the number l of the conjunctive keywords, while the size of the private keys in our scheme is only one element in Z^*_p.
- The size of the ciphertext in the schemes in [8]–[10], [12], and [16] are l + 2, 2l + 2, 2l + 4, l + 1, 2l + 4, respectively, while the size of the ciphertext is l + 3 in our scheme.
- The sizes of the trapdoor in the schemes in [8]–[10] and [12] and ours are linear with the number l of the conjunctive keywords.

Note: It seems that the schemes in [13]–[15], [22]–[25], [28], and [30] have a better efficiency of the ciphertext than our scheme. However, these schemes can only support the single keyword search, while our scheme is more efficient superior than other searchable encryption schemes in the sense of that it allows the conjunctive keyword search.

The computation overhead comparison is also shown in Table III. There are two major time-consuming computation operations: exponentiation and bilinear pairing. The bilinear pairing computation is the most time consuming operation among all computations. It can be found that the proposed Re-dtPECK can achieve a higher efficiency than other PEKS schemes that supports conjunctive keyword search.

B. Efficiency Evaluation by Simulation

We have evaluated the proposed Re-dtPECK scheme by implementing key components on an experimental workbench, including the system global setup, the key generation, the re-encryption key generation, the trapdoor generation and the test algorithms. The pairing-based cryptography (PBC) Library [32] is used. We have selected the type-A elliptic curve parameter, which provides 1024-bit discrete log security strength equivalent to the group order of 160-bit. The exper-iments have been executed on a PC running Windows7 with an Intel core CPU at 2.5GHz and a 4.0 GB of the memory. We have evaluated the communication, the computation and [illegible] to provide such searchable experimen-tal results.

Fig. 4 shows the operation time of each algorithm of our scheme with the different values of l, which is the number of keywords. We have also tested the schemes in [10], [12], [16], and [28]–[30] on the same experimental workbench so that a simulation comparison can be made fair. In order to describe the performance differences, a non-uniform axis has been used in Fig. 4 (c, f, h) to make the distinct values much clearer. Since the schemes in [10], [12], [16], and [28]–[30] do not have $TrapdoorR$ and $TestR$ algorithms, there is only one line in Fig. 4 (k, l). Compared with these schemes, it is easy to find that our scheme almost always has a smaller execution time. In Fig.4(b), our scheme takes 9ms more than other schemes due to that one more group element is randomly selected for generating the public key of the data server. This tiny difference can be virtually ignored because key generation algorithm is executed only once by the TTP.

In our *dPECK* algorithm, the ciphertext is computed as $t = e(X, V)^S$, $C_1 = t \cdot e(g, g)^{-r}$, $C_2 = g^S$, $B_\psi = g^{r \cdot \eta_\psi}$, $0 \le \psi \le l$. Two bilinear pairing operations $e(g, g)$ and $e(X, V)$ can be pre-computed, while they will not consume additional computation time in the document encryption phase. The execution times of the *GlobalSetup*, *KeyGenSer* , KeyGen*Rec*, and *KeyGenTS* algorithms are approximately 9ms, 18ms, 10ms and 9ms. The execution times of the *ReKeyGen* and *TimeSeal* algorithms are about 0.01ms and 19ms. The execution time of *Trapdoor, TrapdoorR* , *Test*, *Test R* , dPECK and Re-dtPECK algorithm grows with the number l of keywords that is extracted from patient's EHR document.

The experimental results indicate that the proposed scheme can achieve a high efficiency, which is desirable for the real world EHR cloud applications.

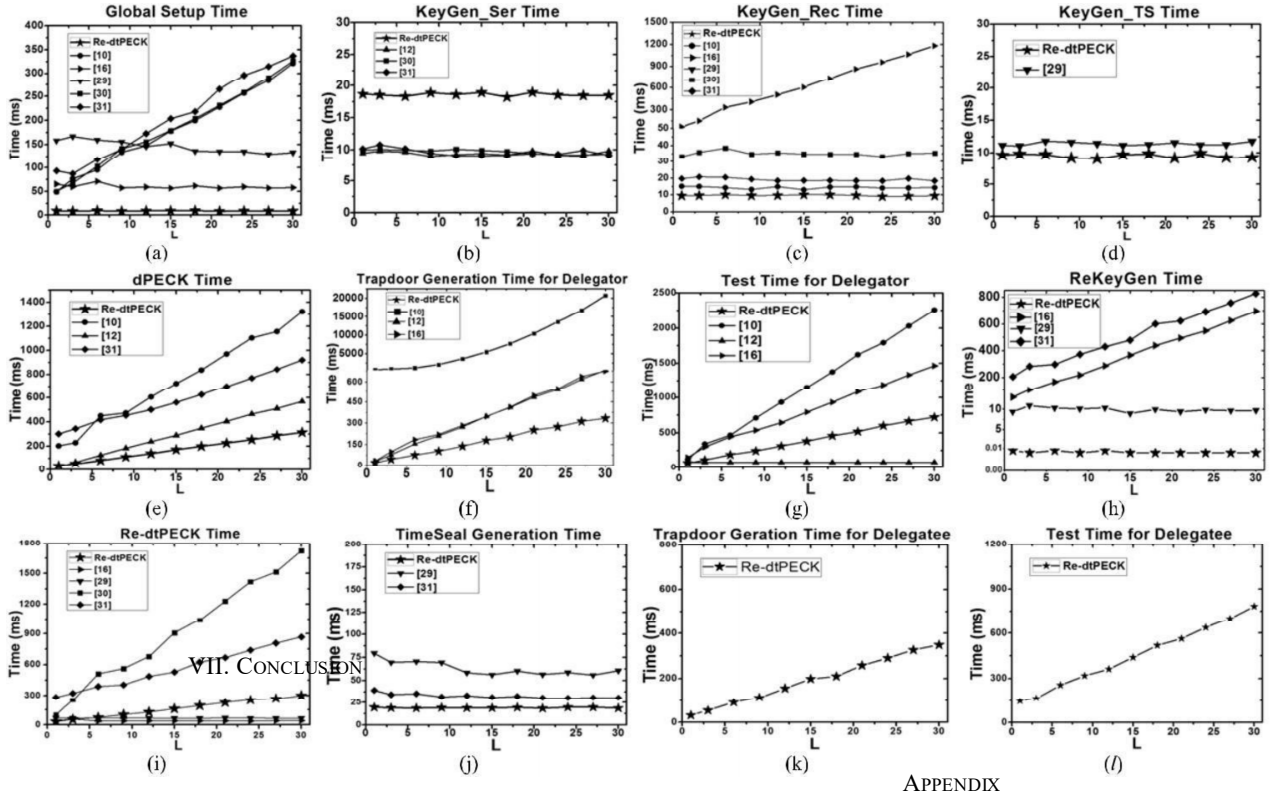


Fig. 4. Time Efficiency of the Scheme. (a) Global Setup Time. (b) Key Generation Time for Server. (c) Key Generation Time for Receiver. (d) Key Generation Time for TS. (e) dPECK Time. (f) Trapdoor Generation Time for Delegator. (g) Test Time for Delegator. (h) Re-encryption Key Generation Time.
(i) Re-dtPECK Time. (j) TimeSeal Generation Time. (k) Trapdoor Generation Time for Delegatee. (*l*) Test Time for Delegatee.

VII. Conclusion

In this paper, we have proposed a novel Re-dtPECK scheme to realize the timing enabled privacy-preserving keyword search mechanism for the EHR cloud storage, which could support the automatic delegation revocation. The experimental results and security analysis indicate that our scheme holds much higher security than the existing solutions with a reasonable overhead for cloud applications. To the best of our knowledge, until now this is the first searchable encryption scheme with the timing enabled proxy re-encryption function and the designated tester for the privacy–preserving HER cloud record storage. The solution could ensure the confidentiality of the EHR and the resistance to the KG attacks. It has also been formally proved secure based on the standard model under the hardness assumption of the truncated decisional *l*-ABDHE problem and the DBDH problem. Compared with other classical searchable encryption schemes, the efficiency analysis shows that our proposed scheme can achieve high computation and storage efficiency besides its higher security. Our simulation results have also shown that the communication and computation overhead of the proposed solution is feasible for any real world application scenarios.

Appendix

A. Security Model of Re-dtPECK

The security model of indistinguishable against keyword guessing attack (IND-KGA) is in accordance with [7] and [22], which is omitted here due to the length limitation. The only difference is that the trapdoor for delegatee is also considered in our proof (Please refer to Appendix. B for the detailed proof).

A Re-dtPECK scheme is IND-CKCTA secure if no probabilistic polynomial time (PPT) attacker $\mathcal{A}$ can win the games below with non-negligible advantage. In the game, $\mathcal{C}$ is the game challenger, k is the security parameter. We consider the following two games.

Game 1: $\mathcal{A}$ is assumed to be a server.

- ***Setup***: The $GlobalSetup$, $KeyGen_{Ser}$, $KeyGen_{Rec}$ and $KeyGen_{TS}$ algorithms are executed. The global parameter GP and keys (skS , pkS), (skR , pkR) $(skTS , pkTS)$ are generated. $(GP, skS , pkS , pkR , pkTS)$ are given to $\mathcal{A}$ while skR is kept secret from $\mathcal{A}$.
- ***Phase 1***: $\mathcal{A}$ makes the following queries.

  (1) Time seal query: $\mathcal{A}$ adaptively queries a time T . $\mathcal{C}$ runs $TimeSeal$ and generates the time seal ST , which is given to $\mathcal{A}$.

  (2) Trapdoor queries: $\mathcal{A}$ adaptively queries a number of keyword sets (denoted as Q) to generate a trapdoor for the user R_i . $\mathcal{C}$ runs $Trapdoor$ and $T_{Q,I}$ is generated and given to $\mathcal{A}$.

  If delegation indicator θ equals to 1, run the queries (3) to (5).

  (3) Trapdoor queries for delegatee: $\mathcal{A}$ adaptively queries several keyword sets (denoted as Q) to generate a

trapdoor for delegatee R_j. C runs $Trapdoor_R$ and TQ, J is generated and given to A.

(4) Re-encryption key queries: A adaptively queries the receivers' ID tuple (i, j), where $i = j$. Challenger C runs $ReKeyGen$ and returns a re-encryption key $rk_{Ri} \rightarrow R_j$ to the adversary.

(5) Re-encryption queries: A adaptively queries (i, j) and the original dPECK ciphertext C_I. Challenger C returns the re-encrypted ciphertext $C_J \leftarrow Re - dtPECK$ to the adversary.

- ***Challenge***: A outputs a challenge set of keywords (W_0^*, W_1^*) and the time (T_0^*, T_1^*). Upon receiving them, the challenger chooses $\sigma, \upsilon \in \{0, 1\}$ uniformly at random and creates the target ciphertext C^* and returns it to A.
- ***Phase 2***: Adversary A is allowed to ask the same types of queries as in phase 1, except the following queries.

  (1) Time seal queries: $TimeSeal(GP, sk_{TS}, T_\upsilon^*, pk_{Ri}, pk_{Rj})$ is not allowed to be queried, where $\upsilon \in \{0, 1\}$.

  (2) Trapdoor queries: $Trapdoor(GP, pk_S, sk_{Ri}, Q)$ is not allowed to be queried if the generated trapdoor is distinguishable for W_0^* and W_1^*.

  (3) Trapdoor queries for delegatee: $Trapdoor_R(GP, pk_S, sk_{Rj}, Q, ST_{\upsilon^*})$ is not allowed to be queried if the generated trapdoor is distinguishable for W_0^* and W_1^*.
- ***Guess***: A outputs its guess σ', $\upsilon \in \{0, 1\}$ and wins the game if $\sigma = \sigma'$, $\upsilon = \upsilon'$.

Game 2: A is assumed to be an outside attacker including the receiver.

- ***Setup***: The $GlobalSetup$, $KeyGen_{Ser}$, $KeyGen_{Rec}$ and $KeyGen_{TS}$ algorithms are executed. The global parameter GP and keys (sk_S, pk_S), (sk_R, pk_R) (sk_{TS}, pk_{TS}) are generated. $(GP, pk_S, pk_R, sk_R, pk_{TS})$ are given to A while sk_S is kept secret from A.
- ***Phase 1***: A makes the following queries.

  (1) Time seal query: A adaptively queries a time T, challenger C runs $TimeSeal$ and generates the time seal S_T, which is given to A.

  If delegation indicator θ equals to 1, run the queries (2) to (3).

  (2) Re-encryption key queries: A adaptively queries the receivers' ID tuple (i, j), where $i = j$. Challenger C runs $ReKeyGen$ and returns the re-encryption key $rk_{Ri} \rightarrow R_j$ to the adversary.

  (3) Re-encryption queries: A adaptively queries (i, j) and the original dPECK ciphertext C_I. Challenger C returns the re-encrypted ciphertext $C_J \leftarrow Re - dtPECK$ to the adversary.
- ***Challenge***: A outputs the target keyword pair (W_0^*, W_1^*) and the time pair (T_0^*, T_1^*). Upon receiving them, the algorithm chooses $\sigma, \upsilon \in \{0, 1\}$ uniformly at random and creates the target ciphertext C^* and returns it to A.
- ***Phase 2***: Adversary A is allowed to ask the same types of queries as in phase 1, except the following queries.

  (1) Time seal queries: $TimeSeal(GP, sk_{TS}, T_\upsilon^*, pk_{Ri}, pk_{Rj})$ is not allowed to be queried, where $\upsilon \in \{0, 1\}$.
- ***Guess***: A outputs its guess σ', $\upsilon \in \{0, 1\}$ and wins the game if $\sigma = \sigma'$, $\upsilon = \upsilon'$.

B. Security Proof of Re-dtPECK

Theorem 1: Suppose the truncated decisional l-ABDHE and DBDH assumption holds, the proposed scheme is IND-CKCTA secure in the standard model. If there is a PPT adversary A, who breaks (t, ε)- IND-CKCTA security of the proposed Re-dtPECK scheme, then we can construct a PPT adversary C to solve the (t_1, ε_1) truncated decisional l-ABDHE and (t_2, ε_2) DBDH problem with

$$\varepsilon_1 \geq \varepsilon,\ t_1 \geq t + t_e[(2l+3)q_{Trapdoor} + (l+5)q_{Re-dtPECK} + 2q_{TimeSeal} + (2l+4)q_{Trapdoor_R}],$$

$$\varepsilon_2 \geq \varepsilon,\ t_2 \geq t + t_e[2q_{TimeSeal} + (l+5)q_{Re-dtPECK}],$$

where t_e denotes the running time of an exponentiation, $q_{Trapdoor}$, $q_{Re-dtPECK}$, $q_{TimeSeal}$, $q_{Trapdoor_R}$ denotes the number of the $Trapdoor$, $Re\text{-}dtPECK$, $TimeSeal$, $Trapdoor_R$ queries.

Game 1: A is assumed to be a server.

chooses $TQ_{i-1}, \varsigma \in Z_p^*$, computes $TQ_{i-2} = g^{\varsigma}$, $TQ_{i,\psi} = g^{m-1} \cdot TQ_{i-1}^{-(y_i)} \cdot \prod_{\mu=1}^{m} H(w_{i,\mu})^{\psi} \cdot X^{\varsigma}$, $0 \leq \psi \leq l$. Since

$$TQ_{i,\psi} = g^{\psi m-1} \cdot T_{Qi,-1}^{-\prod_{\mu=1}^{m} H(w_{i,\mu})}$$
$$X^{\varsigma} = g^{m-1} \cdot TQ_{i-1}^{\alpha^{\psi} \cdot \prod_{\mu=1}^{m} H(w_{i,\mu})^{\psi}} X^{\varsigma_\mu}$$

C has successfully simulated the trapdoor.

If delegation indicator θ equals 1, run the queries (3) to (5).

(3) Trapdoor queries for delegatee: A adaptively queries a keyword set $Q_i = (w_{i,1}, \cdots, w_{i,m})$, $m \leq l$. C randomly chooses $r_T, \varsigma \in Z_p^*$ and sets:

$$TQ_{i,-1} = r_T,\ TQ_{i,-2} = g^{\varsigma},$$
$$TQ_{i,\psi-3} = (g_2 \cdot Y^{-r_T})^{\frac{1}{\tau - H(T, pk_i)}} \prod_{\mu=1}^{m} H(w_{i,\mu})^{\psi},$$

Since $(g_2 \cdot Y^{-r_T})^{\frac{1}{\tau - H(T, pk_i)}} = (Y \cdot g^{-r_T})^{\frac{\alpha}{\tau - H(T, pk_i)}}$, this is a

valid trapdoor for Q_i . Thus, C has successfully simulated the trapdoor $T_{Q_i, J}$ for the delegatee.

(4) Re-encryption key queries: A adaptively queries receivers' ID tuple (i, j), where $i = j$. C runs $KeyGen_{Rec}$ to obtain y_i and y_j . Then, the challenger runs $ReKeyGen$ and returns a re-encryption key $rk_{R_i \rightarrow R_j} = y_i / y_j$ to the adversary.

(5) Re-encryption queries: A adaptively queries (i, j) and the original dPECK ciphertex CI . C runs $ReKeyGen$ and obtains the re-encryption key $rk_{R_i \rightarrow R_j} = y_i / y_j$. Then, the challenger runs the re-encryption algorithm, computes $C3 = (g^{-H(T, pk_i)} pkTS)^{\mu}, C4 = Y_j^{\mu}$, $C5 = g^{\mu}, B_{\psi} = B_{\psi}^{R_i \rightarrow R_j}$ $(0 \le \psi \le l)$ and returns CI' to A.

- ***Challenge***: A outputs a challenge set of keywords (W_0^*, W_1^*) and a target set T_{σ}^*. C randomly chooses $\sigma, \upsilon \in \{0, 1\}$, $W_{\sigma}^* = (w_{\sigma,1}^*, \ldots, w_{\sigma,l}^*)$, and T_{υ}^* is the target. Let $f(x) = x^{l+2}$ and $F(x) = (f(x) - \sum_{\psi=1}^{l} H(w_{\sigma,\psi}^*) x^{\psi}) + 1$. C randomly selects $s \in Z_p^*$ and computes $t^* = e(X, V)^{s*}$. C constructs a polynomial $\Upsilon^*(x) = \eta_l^* x^l + \cdots + \eta_1^* x + \eta_0^*$ such that $x_{\sigma,1}, \cdots, x_{\sigma,l}$ are l roots of the equation $\Upsilon^*(x) = 1$. C sets

$$C_1^* = t^* \cdot Z^{-1} \cdot e(g, g^{F_{\psi} \cdot \alpha_i})^{-1}, \quad C_2^* = g^{s*},$$

$$C_3^* = (g^{-H(T_{\upsilon}^*, pk_i)} pkTS)^{\mu}, \quad C_4^* = Y_j^{\mu}, \quad C_5^* = g^{\mu},$$

$$B_{\psi}^* = (g)^{F(\alpha) \cdot \eta_{\psi}^*}, \quad 0 \le \psi \le l.$$

where F_i is the coefficient of x^i in $F(x)$. Let $r^* = (\log_g g) F(\alpha)$. Since $\log_g g$ is random, r^* is uniformly random. If $Z = e(g_{l+1}, g)$, then

$$C_1^* = t^* \cdot e(g_{l+1}, g)^{-1} \cdot e(g, g^{F_{\psi} \cdot \alpha_i})^{-1}$$
$$= t^* \cdot e(g, g^{l+1} g^{F_{\psi} \cdot \alpha_i})^{-1} = t^* \cdot e(g, g)^{-r^*},$$

$$B_{\psi}^* = (g)^{F(\alpha) \cdot \eta_{\psi}^*} = g^{r^* \cdot \eta_{\psi}^*}, \quad 0 \le \psi \le l.$$

Then, it is a valid encryption for W_{σ}^* and T_{υ}^*.

- ***Phase 2***: Adversary A is allowed to ask the same types of queries as in phase 1, except the following queries.

  (1) Time seal queries: $TimeSeal(GP, sk_{TS}, T_{\upsilon}^*, pk_{R_j})$ is not allowed to be queried, where $\upsilon \in \{0, 1\}$.

  (2) Trapdoor queries: $Trapdoor(GP, pk_S, sk_{R_i}, Q)$ is not allowed to be queried if the generated trapdoor is distinguishable for W_0^* and W_1^*.

  (3) Trapdoor queries for delegatee: $Trapdoor_R(GP, pk_S, sk_{R_j}, Q, S_{T\upsilon^*})$ is not allowed to be queried if the generated trapdoor is distinguishable for W_0^* and W_1^*.

- ***Guess***: A outputs a guess $\sigma, \upsilon \in \{0, 1\}$. If $\sigma = \sigma$, $\upsilon = \upsilon$, then C outputs 1 meaning $Z = e(g_{l+1}, g)$. Otherwise, it outputs 0 meaning $Z = e(g_{l+1}, g)$ but a random element in G_1.

Probability Analysis: In the guess phase, if the adversary is capable to break the scheme with the advantage ε, $e(g_{l+1}, g)$ will appear in the tuple list with probability $1/2 + \varepsilon$ at least. Then, the advantage of C against the truncated decisional l-ABDHE problem is at least $\varepsilon_1 \ge \varepsilon$.

Time Analysis: The execution time of the simulation is dominated by the exponent operation in the query phase. Then

$$t_1 \ge t + t_e [(2l + 3) q_{Trapdoor} + (l + 5) q_{Re-dtPECK} + 2q_{TimeSeal} + (2l + 4) q_{TrapdoorR}]$$

Game 2: A is assumed to be an outside attacker including the receiver.

Proof: Suppose that the attacker A is given a tuple $Tu = (g, g^a, g^b, g^c, Z)$ as an instance for the DBDH problem, where Z is either $e(g, g)^{abc}$ or a random element of G_1. The game between A and C proceeds as follows.

- ***Setup***: Let $X = g^a$, $V = g^b$, the data server's public and private key pair is $pk_S = (V, X)$ and $sk_S = a$. C randomly chooses $y \in Z_p^*$ and calculates $Y = g^y$.
-
-

REFERENCES

[1] J. C. Leventhal, J. A. Cummins, P. H. Schwartz, D. K. Martin, and W. M. Tierney, "Designing a system for patients controlling providers' access to their electronic health records: Organizational and technical challenges," *J. General Internal Med.*, vol. 30, no. 1, pp. 17–24, 2015.

[2] Microsoft. *Microsoft HealthVault*. [Online]. Available: http://www. healthvault.com, accessed May 1, 2015.

[3] Google Inc. *Google Health*. [Online]. Available: https://www. google.com/health, accessed Jan. 1, 2013.

[4] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in *Proc. EUROCRYPT*, vol. 3027. Interlaken, Switzerland, May 2004, pp. 506–522.

[5] Q. Tang, "Public key encryption schemes supporting equality test with authorisation of different granularity," *Int. J. Appl. Cryptogr.*, vol. 2, no. 4, pp. 304–321, 2012.

[6] P. Liu, J. Wang, H. Ma, and H. Nie, "Efficient verifiable public key encryption with keyword search based on KP-ABE," in *Proc. IEEE 9th Int. Conf. Broadband Wireless Comput., Commun. Appl. (BWCCA)*, Nov. 2014, pp. 584–589.

[7] L. Fang, W. Susilo, C. Ge, and J. Wang, "Public key encryption with keyword search secure against keyword guessing attacks without random oracle," *Inf. Sci.*, vol. 238, pp. 221–241, Jul. 2013.

[8] M.-S. Hwang, S.-T. Hsu, and C.-C. Lee, "A new public key encryption with conjunctive field keyword search scheme," *Inf. Technol. Control*, vol. 43, no. 3, pp. 277–288, 2014.

[9] D. Boneh and B. Waters, "Conjunctive, subset, and range queries on encrypted data," in *Proc. 4th Theory Cryptogr. Conf.*, vol. 4392. Amsterdam, The Netherlands, Feb. 2007, pp. 535–554.

[10] B. Zhang and F. Zhang, "An efficient public key encryption with conjunctive-subset keywords search," *J. Netw. Comput. Appl.*, vol. 34, no. 1, pp. 262–267, 2011.

[11] J. W. Byun and D. H. Lee, "On a security model of conjunctive keyword search over encrypted relational database," *J. Syst. Softw.*, vol. 84, no. 8, pp. 1364–1372, 2011.

[12] M. Ding, F. Gao, Z. Jin, and H. Zhang, "An efficient public key encryption with conjunctive keyword search scheme based on pairings," in *Proc. 3rd IEEE Int. Conf. Netw. Infrastruct. Digit. Content (IC-NIDC)*, Beijing, China, Sep. 2012, pp. 526–530.

[13] J. Shao, Z. Cao, X. Liang, and H. Lin, "Proxy re-encryption with keyword search," *Inf. Sci.*, vol. 180, no. 13, pp. 2576–2587, 2010.

[14] W.-C. Yau, R. C.-W. Phan, S.-H. Heng, and B.-M. Goi, "Proxy re-encryption with keyword search: New definitions and algorithms," in *Proc. Int. Conf. Security Technol.*, vol. 122. Jeju Island, Korea, Dec. 2010, pp. 149–160.

[15] L. Fang, W. Susilo, C. Ge, and J. Wang, "Chosen-ciphertext secure anonymous conditional proxy re-encryption with keyword search," *Theoretical Comput. Sci.*, vol. 462, pp. 39–58, Nov. 2012.

[16] X. A. Wang, X. Huang, X. Yang, L. Liu, and X. Wu, "Further observation on proxy re-encryption with keyword search," *J. Syst. Softw.*, vol. 85, no. 3, pp. 643–654, 2012.

[17] R. Canetti, O. Goldreich, and S. Halevi, "The random oracle methodology, revisited," *J. ACM*, vol. 51, no. 4, pp. 557–594, 2004.

[18] M. Bellare, A. Boldyreva, and A. Palacio, "An uninstantiable random-oracle-model scheme for a hybrid-encryption problem," in *Proc. Int. Conf. Theory Appl. Cryptogr. Techn. (EUROCRYPT)*, vol. 3027. Interlaken, Switzerland, May 2004, pp. 171–188.

[19] J. W. Byun, H. S. Rhee, H.-A. Park, and D. H. Lee, "Offline keyword guessing attacks on recent keyword search schemes over encrypted data," in *Proc. 3rd VLDB Workshop Secure Data Manage. (SDM)*, vol. 4165. Seoul, Korea, Sep. 2006, pp. 75–83.

[20] W.-C. Yau, R. C.-W. Phan, S.-H. Heng, and B.-M. Goi, "Keyword guessing attacks on secure searchable public key encryption schemes with a designated tester," *Int. J. Comput. Math.*, vol. 90, no. 12, pp. 2581–2587, 2013.

[21] J. Baek, R. Safavi-Naini, and W. Susilo, "Public key encryption with keyword search revisited," in *Proc. Int. Conf. ICCSA*, vol. 5072. Perugia, Italy, Jun./Jul. 2008, pp. 1249–1259.

[22] L. Guo and W. C. Yau, "Efficient secure-channel free public key encryption with keyword search for EMRs in cloud storage," *J. Med. Syst.*, vol. 39, no. 2, pp. 1–11, 2015.

[23] H. S. Rhee, J. H. Park, W. Susilo, and D. H. Lee, "Trapdoor security in a searchable public-key encryption scheme with a designated tester," *J. Syst. Softw.*, vol. 83, no. 5, pp. 763–771, 2010.

[24] C. Hu and P. Liu, "A secure searchable public key encryption scheme with a designated tester against keyword guessing attacks and its extension," in *Proc. Int. Conf. Adv. Comput. Sci., Envi-ron., Ecoinform., Edu. (CSEE)*, vol. 512. Wuhan, China, Aug. 2011, pp. 131–136.

[25] C. Hu and P. Liu, "An enhanced searchable public key encryption scheme with a designated tester and its extensions," *J. Comput.*, vol. 7, no. 3, pp. 716–723, 2012.

[26] H. S. Rhee, J. H. Park, and D. H. Lee, "Generic construction of designated tester public-key encryption with keyword search," *Inf. Sci.*, vol. 205, pp. 93–109, Nov. 2012.

[27] W.-C. Yau, R. C.-W. Phan, S.-H. Heng, and B.-M. Goi, "Secu-rity models for delegated keyword searching within encrypted contents," *J. Internet Services Appl.*, vol. 3, no. 2, pp. 233–241, 2012.

[28] K. Emura, A. Miyaji, and K. Omote, "A timed-release proxy re-encryption scheme," *IEICE Trans. Fundam. Electron., Commun. Comput. Sci.*, vol. 94, no. 8, pp. 1682–1695, 2011.

[29] Q. Liu, G. Wang, and J. Wu, "Time-based proxy re-encryption scheme for secure data sharing in a cloud environment," *Inf. Sci.*, vol. 258, pp. 355–370, Feb. 2014.

[30] K. Liang, Q. Huang, R. Schlegel, D. S. Wong, and C. Tang, "A condi-tional proxy broadcast re-encryption scheme supporting timed-release," In *Information Security Practice and Experience*. Berlin, Germany: Springer, 2013, pp. 132–146.

[31] J. Li, Y. Shi, and Y. Zhang, "Searchable ciphertext-policy attribute-based encryption with revocation in cloud storage," *Int. J. Commun. Syst.*, doi: 10.1002/dac.2942, 2015.

[32] B. Lynn. *The PBC Library*. [Online]. Available: http://crypto. stanford.edu/pbc/, accessed May 1, 2015.

[33] D. Cash, S. Jarecki, C. Jutla, H. Krawczyk, M.-C. Ro¸su, and M. Steiner, "Highly-scalable searchable symmetric encryption with sup-port for Boolean queries," in *Advances in Cryptology*, Berlin, Germany: Springer, 2013, pp. 353–373.

[34] D. Cash *et al.*, "Dynamic searchable encryption in very-large databases: Data structures and implementation," in *Proc. Netw. Distrib. Syst. Security Symp. (NDSS)*, Feb. 2014, pp. 1–32.

[35] S. Jarecki, C. Jutla, H. Krawczyk, M. Rosu, and M. Steiner, "Outsourced symmetric private information retrieval," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2013, pp. 875–888.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Survey on human Wearable Computing

Dr.S.P.Anandaraj, Professor,
Department of CSE,
Malla Reddy Engineering College for Women
E-mail id:anandsofttech@gmail.com

Dr.N.Kirubakaran
Department of CSE,
Malla Reddy Engineering College for Women
e-mail id:iamkiru70@gmail.com

Abstract:With electronic chips, wearable devices are developed. So many forms of devices available on the market, such as glasses, bracelets that offer exclusive and customized services. These are based on sensor technology. Wearable devices present computer science with new problems and challenges

Keywords: Electronic chips, bracelets, glasses, sensor technologies.

I. INTRODUCTION

A. The Definition of Wearable Devices

A wearable device could be a laptop that's subsumed into the non-public area of a user, controlled by the user, and has each operational and reciprocal constancy, i.e., is often on and always accessible [1]. Wearable devices have a similarcomputing skills as mobile phones and pillcomputers. In some cases, however, wearable devices are additional compete nt for tasks like calculation, navigation, remote image than hand-held devices thanksto theirmovability and characteristics to be careful below.

B. The Development of Wearable Devices

We can have a transparent understanding of the event for wearable devices from [2][3]. Wearable devices have undergone a few years of development since the initial concepts and prototypes appeared within the Sixties. throughout the Sixties to Nineteen Seventies, wearable devices were in their embryonic amount. Individuals designed wearable devices for special functions, interests or events. throughout the amount, wearable devices remained in an exceedingly small-scale field and other peopleseldom understood their roles. In 1966, Edward Thorp, a academician within the Massachusetts Institute of Technology (MIT), fancied a combine of shoes that would be wont to cheat at roulette. this can be the primary wearable device within the world. In 1975, the Hamilton Watch Company launched a "calculator" watch that is that the world's initial carpus calculator. In 1977, the CC Collins designed a wearable device for the blind, that converts pictures captured from a head-mounted camera into tactile grids settled on the blind's vests.

During the Eighties to Nineties, wearable devices entered the first stage of development. individuals began to concentrate to wearable devices. though wearab le technologyhad an excellent improvement, wearable devices were still not sensible for customers and not friendly for users. In 1981, Steve Mann designed a head-mounted camera that to some extent are

ISBN 978-81-939386-1-4

often thought to be the pioneer of Google glasses. within the same year, Steve Mann designed a backpack vogue laptop with text, image and transmission functions, displaying through the helmet. In 1997, Massachusetts Institute of Technology, Carnegie Andrew MellonUniversity, and Georgia Institute of Technology together organized the primary International conference on Intelligent wearablelaptop (ISWC). Since then, sensible wearable computing and sensible wearable devices have attracted wide attention in academeand trade.

Since the twenty first century, wearable devices have entered a complicated stage of development and aroused widespread concern. They become a lot of advanced and ar designed consistent with the requirements of users or the market. Severalcorporations launched severally de signed wearable devices and discharged corresponding code and hardware development platforms. In 2007, James Park and Eric economic expert supported the Fitbit Company that's dedicated to the event of wearabledevices on measuring system and sleep quality detection etc. In 2013, Google launched Google glass and caused a sensation within the world. Meanwhile, Apple, Samsung, Sony and different corporations are developing their sensible watches.

In the next few years, predictably, wearable devices can enter a amount of prosperity. The IMS information [4] disclosed that wearable devic es shipments can reach ninety two.5 million units by 2016. consistent with Juniper's analysis [5], the amount of wearable devices together with sensible watches and glasses can approach a hundred thirty million by 2018. Moreover, consistent with IDC's reports [6], we tend to note that wearable devices had been beneath a fast development; the amount ought to be up to nineteen million by the tip of 2014; and therefore the foreseen shipments together with sensible watches and connected devices can grow at an annual rate of seventy eight and reach 112 million by 2018. Therefore, we can able to believe that wearable devices will bit by bit enter people's lives and convey convenience to human, and wearable market can attract a lot of participants.

At present, there are two standards for classifying wearable devices [7]. One normal relies on product forms, as well as head-mounted (such as glass and helmet), body-dressed (such as coat, underwear, and trousers), hand-worn (such as watch,bracelet, and gloves), and foot-worn (such as shoes and socks).



Figure 1. a range of wearable devices

Another normal relies on product functions, as well as healthy living (such as sport wristband and sensible bracelet), dataconsulting (such as sensible glass and sensible watch), and modality management (such as modality controller). Figure one lists a range of wearable devices.

II. WHAT MAKES WEARABLE DEVICES DIFFERENT?

A wearable device is additional convenient for users to use and carry thanks to its miniaturization, light-weight and dressing. Their functions, forms and usages are completely different from pill computers and mobile

phones. within the wearableconference control in 1997 [8], Bass L. summarized 5 characteristics that a wearable device ought to have [9]: 1) It is also used whereas the user is in motion; 2) It is also used whereas one or each hands are free or occupied with tasks; 3) It exists at intervals the corporeal envelope of the user, i.e., it ought to be not simply hooked up to the body however becomes AN integral a part of the person's clothing; 4) It should permit the user to keep up control; 5) It should exhibit constancy, within the sense that it ought to be perpetually obtainable. Steve Mann additionally provided the definition of wearable devices and delineate them from 3 operational modes and 6 attributes at the International Conference on wearable Computing (ICWC) [1] control in 1998, The 3 operational modes embrace constancy, augmentation and mediation. and therefore the Six attributes embrace unmonopolizing of the user's attention, emancipative to the user, noticeable by the user, manageable by the user, alert to the atmosphere and communicative to others.

However, wearable devices show additional options as they evolve, like diversity and concealment [10]. Wearable devices not solely amendment the human-computer relationship and therefore the method folks use computers. Moreover, they manufacturea major influence on people's life and work.

III. software system FOR wearable DEVICES

The technology of wearable devices isn't mature. the event of wearable devices is therefore guaranteed to encounter variedissues like purposeful singlene ss, incompatibility between software and software system, convenience of human-computer interaction, information transmission, confidentiality of the data, energy consumption problems brought by continuous running. This chapter discusses some primary analysis problems within the development methodof wearable devices.

A. operative System

Operating system is that the interface of hardware and software system. Its perform is to manage hardware, software system and information resources, to regulate program execution, to boost human-computer interaction, to change users to own a decentoperating setting, and to supply services for users and support for different applications.

The software on wearable devices has seasoned years of development [11]. As early as in 2000, IBM collaborates with the knownJapanese watch manufacturer national to launch a wise watch named WatchPad with Linux as its software. Fossil designed in 2003 a carpus device known as carpus personal organizer. It equipped with PalmOS [12] software and supported screen bit, that were very fashionable at that point. Additionally, Microsoft designed in 2004 the SPOT system for good watches. In 2013, Samsung discharged its 1st good watch Galaxy Gear mistreatment robot as software. After that, Samsung launched the second generation of good watch running Samsung's severally designed software Tizen. In March 2014, supported robot, Google launched a wise watch dedicated software known as robot Wear [13], whose operation is enforced through Google Now's voice commands. robot Wear is anticipated to make commonplacean identical and standard softwareplatform, fast the event of wearable devices.

At present, there exist a range of operative systems in wearable market. however they'll not be convenient for users to use. Developers ar tough to decide on that software for the device. and therefore the application for one software isn't appropriatefor one more.

Since software is crucial for wearable devices, we must always style wearable operative systems by

taking the options of wearable devices into consideration in order that we will reach the subsequent objectives.

1) Convenience: the planning of the software ought to be a lot of convenient for users to use wearable devices.
2) Effectiveness: The software ought to be managed a lot of effectively and cash in of resources like hardware, software systemand information of wearable devices.

3) Scalability: The software ought to allow new system functions to be developed, tested and enclosed.

4) Openness: The software ought to support integrated and cooperative network work of various makers and devices in order that it can do the movability and ability of applications.

5) Multitasking: The software ought to be ready to run multiple applications at the same time.

In general, there ar many choices for developers to develop wearable software [14][15]:

1) evelop any on the idea of palm software used on mobile phones or different terminals. The advantage of this approach is to shorten the event cycle and to utilize existing applications. Its disadvantage is that there exist some issues on movement the prevailing applications.
2) Develop proprietary software supported UNIX system. This approach has sturdy applicability. However, it's troublesome for development and establishing the corresponding software system of ecological surroundings.

3) Develop web-based software [14]. This approach would take full advantage of resources and support of the net server, regulate and structure software dynamically in keeping with the particular needs, minimize the resource needs to wearable devices. The approach has nice potential for development, however can take a protracted time and an oversized value.

B. management System

Wearable devices run unceasingly, and are invariably able to act with users. they'll collect user data the maximum amount as potential, and store them in information. The user will access the information to question necessary data via the management system. As an example, the help good cane designed by Eagle Ugintaite [7], utilizes intrinsic sensors to record user's pulse, pressure level and temperature in real time, and displays health information through its liquid crystal display screen. The user will consult the information to get varied information concerning his/her recent health standing.

In addition to the essential data as well as gender, age, height and weight, wearable devices also can record user's health standing, consumption habits, personal temper, and preference data to paint and food, etc. These data could also be helpful for users achieving their goals. as an example, once it's time for lunch, wearable devices are ready to search close restaurants meeting the user's style in keeping with the information data, and supply hints to the user; once looking, the devices also can suggest bound commodities that the user may have or could also be fascinated by supported personal preference; once tying up good appliances, wearable devices will s end personal data keep within the information to the terminals, and instruct good appliance system to

regulate air con in real time in keeping with the user's temperature, or mechanically open the audio device and play music fitting the user's mood at that point.

At present, the key task of the management system for wearable devices is simply to trot out easy information, even some devices don't embrace the perform of management. The response speed and process power might not essentially meet the requirements of devices.
Due to the various functionalities of wearable devices, the recorded information are kaleidoscopically and should embrace the essential personal data, physical health information, and external surroundings dynamical informatio n etc. To store this information, not solely native databases, however conjoi ntly cloud databases are needed and a reliable management system becomes necessary. Therefore, developers ought to style specialized management systems that are ready to manage and operate varied information, being light-weight and having quick response speed; or ought to transfer existing management systems on mobile phones to wearable devices.

c. Network Communication Protocol

Wearable devices may have to exchange information with alternative devices lik e mobile phones, computers and alternativewearable devices. Therefore, network protocols for wearable devices become necessary. Network protocols outline the network communication mode of wearable devices,
and verify exchange info and issues associated with synchronization. Bluetooth (IEEE 802.15.1), ZigBee (IEEE802.15.4), WiFi, NFC (IEEE 802.11) ar four presently standard short-range wireless communication protocol standards [16]. In 2013, Broadcom launched Wireless web property for Embedded Devices (WICED) [17]. WICED simplifies the implementation of WiFi in a very series of client physics. It are able to do WiFi networking application or connect mobile phones, tablets computers and wearable devices for information sharing. Bluetooth good [18] is AN improved version of Bluetooth four.0 [19] with the facility reduced by 10 to twenty times. what is more, even though the radio is enclosedmost of the time, devices will still keep connected; once information ar prepared, devices is woke up like a shot inside a awfullyshort time. during this manner, battery life is extended, that makes Bluetooth good a lot of appropriate for wearable devices. At present, network communication protocolsfor wearable devices are comparatively easy and focus totally on wireless functions. However, wearable devices can eventually implement a lot of and a lot of functions like those enforced in mobile phones, likeWAP, GRS, GPRS, massive fileor information transmission, etc. Therefore, a lot of reliable network communication protocol supports are demanded.

Developers ought to transfer network communication protocols running on pill computers or mobile phones to wearable devices, or style special network communication protocols for wearable devices, that are a lot of energy-efficient, safe, and with high outturn.

D.Application Development Platform

The fast development of mobile phones attributes the success to the big range of accessible applications. Similarly, the dearth of applications is one amongst the obstacles for

ISBN 978-81-939386-1-4

the fast development of wearable devices. computer code development needsdevelopment environments and an entire set of tools, together with modeling tools, interface tools, project management, configuration management tools, and testing tools, etc. computer code development kit is commonly accustomed support a specific computer code engineering methodology, cut back the burden of manual management, and build computer code a lot of systematic
like alternative computercode engineering strategies. At present, there ar several hardware style platforms for wearable devices. In 2013, Broadcom launched a development platform known as WICED [17]. Bluetooth, WiFi, NFC and positioning technology is integrated into wearable devices. The wireless networking perform with low power consumption, high performance and ability can even be embedded into devices. because therepresentative
ofthe wearable systemplatform supported ARM desig n, Freescale launched the ASCII text
file, scalable wearable Reference Platform (WaRP) [20] with over fifteenmanufacturer collaborators together with Kynetics, RevolutionRobotics, Circuitco. The platform supports openoperating system like mechanical man and UNIX operating system and
has options like quantifiability and adaptability.Wearable market is presently within the stage of development and promotion of hardware, whereas application development is insulating material behind. so as to accelerate the software package application development of wearable market, numeroussoftware package reference platforms are bobbing up. In 2014, Google discharged the software package development kit (SDK) as well as associate degree mortal and alternative tools for mechanical man Wear [21], which may enable developers to integrate mechanical man Wear platform with their own devices and applications that
may be accessible to transfer. Developers of the mechanical man Wear ecological circle producing will utilize all the tools they have to start out creating apps for the new devices. the event of SDK for mechanical man Wear is anticipated to bring richer APP expertise for future good watches. Tizen is associate degree open supply OS supported UNIX operating system for mobile phones
and alternative good devices. it's a whole development platform, as well as machine, IDE etc. At present, Samsung's Galaxy Gear2 good watch is running on this OS. If the promotion
is wide successful , we tend to believe that Tizen can contribute to the
event of wearable devices.

However, it's still lack of mature software package development platform
for wearable devices. a vital issue for developers is to decide on that platform to develop application software package. additionally, developers cannot
verify that applications aredeveloped and accessible. Therefore, developers have many choices [22]:

1) Develop application software package that supports specific operative systems for one platform. This approach would changedevelopers' work, however the ensuing application might not be ready to run on wearable devices with alternative completely different operative systems;
2) Develop native applications for every platform, however the technology and value of application maintenance for every platform become the massive challenges;

3) Develop
mobile internet applications, therefore on cut back the native code for every platform. however the applying might not be ready to meet the market demand.

E. Privacy and Security

Wearable devices will collect period of time user data therefore on offer effective customize

d services for users.
These knowledge contain numerous user informations, for instance, geographical location, living habits, temperature, heart rate, account watchword and spoken language, etc. once mishandled, it should bring nice danger to the user's privacy and security, and hurtthe user's property or maybe personal safety. With the wide use of wearable devices and also the rapid climb of applications, security and privacy problems ought to become additional and additional vital. The communication of wearable devices is principally supported wireless network. Thus, additional non-public data could also be simply attacked or taken.

Wearable business isn't mature. there's still lack of a whole and effective program to guard privacy and security. whereas the business might neglect security issues because of price problems, researchers operating in wireless sensors and mobile application have done tons of basic work which will facilitate to unravel the privacy and security issues in wearable devices. for instance, Liu et al. studied the protection issues on mobile application [23]. Ameen et al. analyzed the issuesof information security and privacy protection on wireless detector network within the health care field [24].

We can think about from the subsequent aspects to guard user's data and privacy:
1) Research on reliable network communication protocols for wearable devices thus on make sure the security of information throughout transmission;

2) The system ought to have permission setting. wearable applications will solely be allowed to get necessary information thus on cut backinformation exposure;

3) Reasonable code patterns ought to be projected to unravel these issues. for instance, loosing the binding between non-public information and real names and intermixture numerous could facilitate to guard user's privacy and security.

F.Energy Consumption

Because wearable devices will solely usebattery, inste ad of stationary power, as their power provide, it's a lot of tedious to charge wearable devices than mobile phones. Frequent charging or substitution battery will inevitably cut back the practicableness of devices and also the preference and satisfaction of users. additionally, the massive quantity of energy consumption for devices will turn out nice heat. If the cooling drawback can't behandled properly, it'll injury user expertise or perhaps cause temperature scald. Therefore, the energy consumption of wearable devices is a difficultywarrant attention.

At present, makers management energy consumption of wearable devices or mobile phones through the planning of hardware or operative systems. Since the start of victimization wearable devices, their battery life isn't terribly satisfactory. Taking sensible watches as associate example, the battery lifetime of Moto thirty six0 is up to sixty hours [25] and also the one amongst LG G Watch is merely 36 hours [26]. Energy consumption management is a vital issue for wearable devices.

In addition to hardware and operative systems, we are able to additionally think about up energy consumption management from a high-level application layer.
Particularly, the management of energy consumption may be thought-about from the subsequent aspects at the appliance layer.

1) Reduce hardware electricity consumption through affordable invoke of system genus Apis. for instance, Hao et al. and Li et al. projected the code-level energy consumption analysis strategies on mobile applications [27][28]. Such strategies will be extended to wearable devices to achieve the purpose

of reducing energy consumption through invoking less-energy-consumption API sing affordable invoking sequence.
2) Create adaptative energy-sensitive applications to regulate mechanically energy usage. once the energy is decent, prime quality services are going to be provided; otherwise, unimportant applications are going to be turned off so as to extend the usage time. Mizouni R et al utilized in [29] such adaptative strategy to cut back the energy consumption of sensible phones in mobile applications.

3) Adopt load-balancing technique to transfer advanced calculation to the mobile terminal via wireless communication network, so reducing wearabledevice's own energy consumption. Kwon et al. bestowed a way to unravel the matter [30]. an identical approach may be introduced into wearable devices through substitution high calculation energy consumption with low communication energy consumption.

G. Human-Computer Interaction

A major feature of wearable devices is to gather user data, understand user physical conditions and therefore the modification of external surroundings, complete numerous commands, and assist or cue the user mechanically [14]. wearable devices pursue people-oriented, specificallyrequiring wearable devices being a lot of appropriate for the user. wearable devices would be best to understand, acknowledge and perceive human feeli ng, and provides intelligent, sensitive, friendly response. Human-machine interaction could be a key technology of wearable systems, that ought tosolve the interaction between users and wearable devices and improve the power of environmental awareness. Therefore, advanced interaction means that square measure a hotspot within the analysis of wearable devices.
There square measure many ways for users to move with wearable devices [14]:

1) Contextual Awareness: wearable devices incessantly run and collect knowledge, however in most cases, the user doesn't use them. wearabledevices ought to run severally, understand the external surroundings, and transfer helpful data to the user. Starner T et al. planned in visual surroundings perception technique for wearable devices, and distinguished that a wearable device will observe its user to supply lucky data, manage interruptions and tasks, and predict future wants while not being directly commanded by the user.

2) Augmented Reality: that's a technology that enhances users' awareness to universe through the data like sound, video, graphics or GPS knowledgegenerated by the pc.

The goal is to use virtual data to the important world and to put virtual object, scene or system message generated by computers to the importantscene. Zhou et al. bestowed a style approach and a series of sensible proposals of wearable user interfaces in real increased surroundings.
3) Non-keyboard input: the foremost acquainted manner for folks to input data into computers or mobile phones is thru keyboard or mouse. but it's not possible to attach such input devices to wearable devices due to their miniaturisation and light-weight.
Users will move with wearable devices through non-keyboard ways in which like voice, handwriting, gestures, knowledge glove etc. For the disabled, the interaction ways in which of ancientgood devices cannot bring traditional expertise. however they will wear a wearable device that receives messages from alternative sensors, thatsquare measure transmitted to their sensory system once analysis. for instance, for someone whose tympanum is broken, hearing

devices square measure directly connected to his/her bone, that permits him/her to sense the voice that's not passing through ears.

Therefore, compared to computers and mobile phones, wearable devices will offer many alternative ways in which of human-computer interaction for users to strengthen their expertise. however these ways in which cannot meet the demand of all users, developers ought to strengthen the study of human-computer interaction technologies:

1) Apply existing mature human-computer interaction ways in which to wearable devices, like handwriting, voice and alternative non-keyboard input. These ways in which will be simply complete in wearable devices and higher accepted by users.
2) Strengthen the analysis of currently-immature human-computer interaction ways in which, like discourse awareness, increased reality, me diatereality, etc. These ways in which will enhance user expertise and build users have larger interests in wearable devices.
3) Propose new human-computer interaction ways in which. Some ways in which is also appropriate for specific teams or specific surround ings. However, this approach can increase the burden to developers, and what is more, it's going to take time to review and observe these new ways in which.

H. package Engineering

With the large quality of wearable devices, package engineering for wearable devices is turning into a lot of and a lot of necessary. though the analysis of package engineeri ng for wearable devices continues to be rare, we are able to predict some aspects/issues which will be of import and price studying:

1) Demand analysis: a way to excellent the necessity documents ought to be a drag price finding out. From users' comments on the applying, the discussion within the BBS or the analysis of connected on-line articles, users' expectation for the functionalities of the wearable devices will benoninheritable and therefore be accustomed impro ve the necessity documents.

2) Code recommendation: several functions square measure probably to be reused in several applications. Through analyzing ASCII text file of existing applications, we are able to suggest operate codes to developers to attain quick development.

3) Application transplantation: it's not necessary to style completely different applications for every OS. we are able to produce API mapping among completely different platforms, and develop a perfect compiler that may complete the applying development for multiple platforms. Software engineering for wearable devices is associate rising field that gives several opportunities for researchers to analyze to.

I. Big Data

With the event of science and technology, we've entered the time of "big data" with "4V" , specifically the amount, the speed, the variability, and therefore the truthfulness. net technology and mobile applications promote the event of the large knowledge. Laurila et al. mentioned the opportunities and challenges brought by massive knowledge in mobile applications . Big knowledge technology can play a good role in promoting the event of wearable trade. an outsized quantity of information are generated in mistreatment wearable devices, as well as basic personal data, health standing, moreover because the preference to food, fabric and color etc. we are able to gain tons of helpful data by mistreatment massive k nowledge analysis technology, which can offer a good facilitate for research project, social development and users' life. massive knowledge technology will

be accustomed analyze the massive amounts of information collected by wearable devices, and to sight user's body health factors. for instance, the analysis for pulse and force per unit area will perceive user's body standing and potential risks. it's associate undeniable fact that massive knowledge technology will facilitate wearable-device users get pleasure from a lot of convenient life. Redmond et al. mentioned the importance of huge knowledge technology for wearable devices in health care .

Conversely, wearable trade also will promote the event of huge knowledge. Compared to mobile phones, wearable devices can turn out larger quantity of information. So itneeds massive knowledge technology to subsume these collected knowledge even a lot of. for instance, the wearable market has emerged numerous good bracelets with advanced functions, style and parameters; but, these wearable devices cannot use massive knowledge technology for process these knowledge to supply helpful data to users.

Wearable devices and large knowledge technology can complement one another and pursue a standard development. within the massive knowledgesurroundings, wearable market can have nice development, produce monumental we alth, and build people's life a lot of convenient.

IV. CONCLUSION

Wearable devices can become the thought of the event of good devices, and dramatically modification fashionable manner of life. Currently, the eventof wearable devices continues to be in its immature stage, and therefore the major functions specialise in running calculation, navigation, remote image and alternative connected services. However, these services also can be achieved on Gadgets. Meanwhile, analysis on hardware materials and battery life has not achieved a breakthrough; restricted screen house makes the merchandise style terribly difficult; application package development continues to be within the initial stage. because of these issues, it'll take an extended time for wearable devices to become the thought of market.

References

[1] S
[1] The History of Wearable Tech, From the Casino to the Consumer. http://mashable.com/2014/05/13/wearable-technology-history/,Accessed on 18 October 2014.
[2] The History Of Wearable Technology. http://www.redorbit.com/education/reference_library/general-2/history-of/1113176835/the-history-of-wearable-technology/, Accessed on 18 October 2014.
[3] Wearable Technology Market Suited for Rapid Growth. http:// press.ihs.com/press-release/design-supply-chain/wearable-technology-market-suited-rapid-growth, Accessed on 18 October 2014.
[4] Press Release: Mobile Smart Wearable Device Shipments to Approach 130 Million by 2018, Juniper Research finds. http:// www.juniperresearch.com/viewpressrelease.php?pr=414, Accessed on 18 October 2014.
[5] Wearables market to take off, hit 112M devices in 2018. http://www.computerworld.com/article/2488133/emerging-technology/wearables-market-to-take-off--hit-112m-devices-in-2018.html, Accessed on 18 October 2014.
[6] G. Chen. Smart Wearable Change the World: The Next Business Tides. Benjing: Publishing House of Electronics Industry, 2014.
[7] 1998 International Conference on Wearable Computing. http:// wearcam.org/icwc/index.html
[8] D. Y. Chen. The Evolution and Trend of Wearable Computer (1). Journal of Chongqing University (Natural Science Edition), 2000, 23(3): 119-124.
[9] From IBM to Microsoft - A Brief History of The Smartwatch. http://www.ibtimes.co.uk/smartwatch-history-apple-iwatch-samsung-galaxy-gear-503752, Accessed on 18 October 2014.
[10] Palm OS. http://en.wikipedia.org/wiki/Palm_OS, Accessed on 18 October 2014.

[11] What is Android Wear. http://www.techradar.com/news/portable-devices/google-android-wear-what-you-need-to-know-1235025#articleContent, Accessed on 18 October 2014.

[12] D. Y. Chen. The Evolution and Trend of Wearable Computer (2). Journal of Chongqing University (Natural Science Edition), 2000, 23(4): 142-148.

[13] Wearable Reference Platform. http://cache.freescale.com/files/32bit /doc/fact_sheet/WARPFS.pdf, Accessed on 18 October 2014.

[14] Y. Liu and D. Hatzinakos. Human acoustic fingerprints: A novel biometric modality for mobile security. Acoustics, Speech and Signal Processing (ICASSP 2014), 2014 IEEE International Conference on. IEEE, 2014: 3784-3788.

[15] M. A. Ameen, J. Liu, and K. Kwak. Security and privacy issues in wireless sensor networks for healthcare applications. Journal of medical systems, 2012, 36(1): 93-101.

[16] Y. Zhou, B. David, and R. Chalon. Innovative user interfaces for wearable computers in real augmented environment. Human-Computer Interaction. Interaction Techniques and Environments. Springer Berlin Heidelberg, 2011: 500-509.

[17] Big Data. http://en.wikipedia.org/wiki/Big_data, Accessed on 18 October 2014.

[18] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bomet, T. M. T. Do, et al. The mobile data challenge: Big data for mobile computing research. Pervasive Computing. 2012 (EPFL-CONF-192489).

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Classify the Peer-To-Peer Networks Videos and Video Servers Using AODE Algorithm

M. Narayanan
Professor, Department of Computer Science and Engineering
Malla Reddy College of Engineering
Maisamaguda, Dhulapally Post, Secunderabad, Hyderabad, Telangana, India
narayanan_baba@yahoo.com

***Abstract*: As one of the wildest emerging technologies P2P has been enticing attention in the field of live streaming and VoD (Video-on-Demand).Video plays an energetic part for communication and any kind of relaxation activity for entertainment. In order to provide reasonable service across all seasons, two machine learning techniques are used wherein availability of server depends on ratios of the season's hits. Popular videos are sorted out based on most number of hits initially, the retrieve phase chooses one or more related cases from the preceding Popularity videos that are stored. The updated / modified video records are reused as per query. In the revise phase, the present popularity record is updated. Finally, updated popularity records are preserved in the retain phase. Seasonal and non-seasonal video server is grouped by applying AODE algorithm. The content of the video categorized on the basis of listener's test at the commencing is additionally examined emerging in 90% clarity of classification.**

Keywords: Averaged One-Dependence Estimators, Case-Based Reasoning, Categorizing Video Server, Peer-to-Peer Networks, Sessional Based Videos, Video-on-Demand,

I. INTRODUCTION

Video is believed as a kind of amusement that provides the vision of steady movement of a report across the portrayal of a sequence of pictures and is recognized as a movie. Until 19th century, the world couldn't unravel its mystery though the origin dates back to 2nd century in China. Slowly, television became the predominant source of entertainment, replacing newspaper and radio. One of the reasons for its popularity above newspaper and wireless is the discernible charm that enables viewers to investigate deeper and be engrossed in the experience.

Towards the conclude of the 20th century, advancing pictures appeared alongside the progress of the Internet and World Wide Web. The two foremost services endowed via the Internet are Web browsing and file transfer. But, requirement of the clients is no longer fulfilled with these services offering information about text, images and document exchange.

The inspiration behind the research for finding ways of conveying live media over the Internet to a personal computer is the success of radio and television broadcasting. Consequently, multimedia data like sound and video above the web are transmitted. Like each supplementary usual text and executable files, the whole multimedia contents were distributed in the alike and sent as files by employing file downloading protocols. The finished file transfer in the download form is spread on the basis of variations in the size of the mass media file and the bandwidth of the channel across that it is carried.

Case-Based Reasoning (CBR) algorithm is utilized for categorizing well recognized videos and Averaged One-Dependence Estimators (AODE) algorithm is utilized for categorizing video server into seasonal and non-seasonal. Literature review of P2P networks is included in Section 2, followed by proposed work in section 3. Implementation is explained in section 4 and finally section 5 comprises of the conclusion.

II. LITERATURE SURVEY

In overlay network distribution system, P2P (Peer-to-Peer) computing is the most sought-after. In P2P systems, super peer network depicts a modern centralized topology design. Hence this overlay helps in enhancing the presentation of P2P requests, exceptionally live streaming. The problem of node failure is addressed in this work by applying gossip protocol based on Firefly algorithm. From a group of peers, it selects the fail over node in super peer network to maintain the overall performance of the network. In order to achieve this, a PeerSim simulator was used to simulate this network model [1], [2].

P2P Networks presenting services on the basis of seasonal and non-seasonal video is the main focus of this work. Case-Based Reasoning algorithm is utilized for categorizing acquainted videos and Averaged One-Dependence Estimators (AODE) algorithm is utilized to categorize video server into seasonal and non-seasonal. Adult and regional domains are classified as non-seasonal with maximum hits in every season. On this basis, the video service forwards specific server for providing service to clients. Hence, by using this type of algorithms, the P2P network offers quality service during all seasons and depending upon the hits of the season the server is made available [3].

Management of video contents in P2P computing is discussed in this paper. Usually accepted data segments are allocated on the cache. Cache replacement algorithms replace old data segments by new data segments. Because of unavailability or lower presence of old data segments, the data contents are less found. Multilevel cache addresses this issue by protecting the old data segments and hence the speed is increased. Clusters based on bandwidth are suggested to deal with connectivity issues. [4].

Video streaming is taken into account to be one amongst the foremost common application of P2P computing and close peers facilitates for accessing information segments. Cache containing popular data segments is downloaded from a video server or a neighboring peer. New data segments replace old data segments by cache replacement algorithm. To circumvent the defeat of old data segments nevertheless being accepted the believed of multi-level cache is used. This maintains the old data and increases the rate of success. This algorithm efficiently grips the data segments [5].

To examine P2P traffic and to find a competent resolution to tackle the alike employing Unsupervised Machine Learning way is the target of this paper. An expectation–maximization (EM) algorithm is counselled to become the maximum probable estimation of peers. This is a type of artificial intelligence technique. The EM successive approximation will trade between an expectation (E) step and a maximization (M) step. Using the current estimation for the factors, Expectation (E) step will create a function for the expectation of the log possibility calculation. Alternatively, maximization (M) step estimates the factors maximizing the feasible log-possibility found on the E step [6].

P2P computing has its own challenges like offering quality service mainly due to network traffic issues. For a reducing traffic, this paper focusses on studying traffic by engaging adaptive routing and design observation. The traffic depends upon hidden listening peers in traffic statistics. To get the highest likelihood approximation of peers, an iterative technique called adaptive routing scheduling algorithm is suggested. It consists of two parts; using the Path Allocation Request (PAR) the first part gives the next peer location. The second part explains about the exact response given to the node by the peer in Explicit Adaption Request (EAR) [7].

This study reveals the introduction of mesh based Peer-to-Peer Networks to avoid overload problem in servers. Downloading data might be a trial after colossal numbers of users endeavor to admission the server. This mesh network overcomes the issue since there is no master/slave relationship among the clients and is apt for live streaming. The demanded data from the server is stored in the cache of the peers. New data segments are stored in the head part and the old ones in the tail part. By synchronizing the tail dispense with neighboring peers, inaccessibility of information at the peers are often addressed [8]. In competitive scrutiny, analogy amid an online algorithm and an optimal offline algorithm is done. The piece selection subjects that external after streaming layered video content above peer-to-peer webs are introduced. This algorithm periodically decides the pieces that are to be requested from other peers for downloading [9].

The protocol and algorithm suggested by Lloret et al (2012) is methodical to enable selecting the best relevant video characteristics and highest coding methods to hold the video streams. A video controller server decides the best video server. Shown as the algorithm presentation and the web presentation are selected by the video controller server [10]. The results are judged against those available in the journalism and recommend areas for upcoming work on network supervising, P2P applications, and network design [11]

To tackle the P2P layered video streaming, Hua et al (2013) gives an illustration about a novel block scheduling scheme. The main function is about striking a good balance between the different aspects. The block arranging subject is next adjusted to an optimization subject that uses the priority sum of the distributed video blocks. Simulation is grasped out to assess the presentation of the algorithms. The results display that the suggested algorithm is successful in terms of bandwidth utilization and video quality [12].

The counselled method is not probable to be utilized for green certification or legislation/regulation satisfaction because of its nature as an estimate. Further, the system is not meant for developing new products. However, the process can boost up monotonous product development techniques by providing immediate and satisfactory accurate evaluation method [13].

Context-aware peer invention and peer association are counselled on the basis of the request context information. The advantages of CA-P2P in achieving fast discovery and association have been demonstrated through simulations. CA-P2P is promising to support existing and emerging proximity-based services, such as commercial advertisement and emergency/disaster communication [14].

III. PROPOSED MODEL

A. Categorizing the Popular Video

Case-based Reasoning is an area of artificial intelligence which comprises of the theoretical foundations analysis, system growth and realistic application built on experience-based problem solving. At present, the CBR is believed as one of the most enthusing and core scrutiny areas. It has the scope of generating diverse methods and applications and then efficiently apply them for determining the problems in the specific area.

One or several appropriate cases (made earlier/stored experiences) are chosen to rectify the

new complications that arises. The answers/responses from the cases are perused here. The key assumption in this approach is that similar difficulties have similar solutions/remedy. Subsequent to selecting the alike kind of cases, the answer(s) from the preceding case(s) are adjusted to coil into a resolution for the present problem. This can be stored in the case-base to enhance its capability, thus implementing a learning behaviour.

B. Categorizing Video Server using Averaged One-Dependence Estimators

It is a probabilistic classification learning algorithm utilized for the purpose of categorization. During the sequence of training, a three dimensional combined frequency table is generated wherein the benefits of the class indexes one dimension and the supplementary two dimensions are accompanied by the benefits of the characteristics. This system was built for dealing with attribute independence setback of the well-known naïve Bayes classifier. This algorithm differs strikingly from the naive Bayes because of the fact that it builds more precise classifiers by slightly enhancing the amount of calculation.

Averaged One-Dependence Estimators (AODE) possesses obtained a lot of attention in new of the enhanced efficiency in training and classification. The result and discussion section encompasses the different classes of video server and how the proposed video system categorizes video server into Seasonal based video server and non-seasonal based video server on their high classification efficiency.

Every seasonal complete one-dependence estimators in AODE is learned as also the type of video request received from the video server. The type of request could be BASICS OF COMPUTER, C++, ORACLE, NS2, VB, MATLAB, VB.NET, LINUX, JAVA, or others. And the forecasting is created by averaging the forecasting of the entire season using one-dependence estimator. A one-dependence estimator is crafted for every single attribute for cutting its intricacy and averting construction learning. The attribute is the basis of all other characteristics. Mentioned below is the AODE classifier.

$$P(S|F) = \frac{\sum(Seasonal(Ai) \geq m\ P(S,F) \prod_{j=1}^{n} P(Fj|S,Fi)}{\sum(S \in F \sum(Seasonal(Ai) \geq m\ P(S,F) \prod_{j=1}^{n} P(Fj|S,Fi)}$$

Where,

P : Probability
Ai : Set of Training Instances
m : Threshold
Seasonal: Defined as the seasonal BASICS OF COMPUTER, C++, ORACLE, NS2, VB, MATLAB, VB.NET, LINUX, JAVA, or others
S : Video Server
F : Features ∈ Seasonal

The restricted possibility is P (S|F), the degree of belief in S, given that the proposal F is right. That is the theory determining the probability of an incident given that another incident has happened either by hypothesis, affirmation or confirmation. The video server locates the plethora of seasonal videos which may be BASICS OF COMPUTER, C++, ORACLE, NS2, VB, MATLAB, VB.NET, LINUX, JAVA or others. Compared to representative discovering algorithms, the examinations clarify that AODE accomplishes enhanced probability estimation. For enhanced performance of probability estimation and managing bigger list, the AODE must be improved. The drawbacks in a feature assortment of AODE for probability estimation is reviewed and examined in order to suggest an efficient algorithm.

IV. RESULT AND IMPLEMENTATION

The development handles the categorization of video server and its content. AODE categorizes Video Server into seasonal and non-seasonal server. Initially, the video content is categorized on the basis of the hypothesis which could be BASICS OF COMPUTER, C++, ORACLE, NS2, VB, MATLAB, VB.NET, LINUX and JAVA

By and large, the number of servers for classification is 200 with 272 videos in a plethora of set-up such as FFmpeg, mp4, DVD, avi and mjpeg. The seasonal and non-seasonal server is first calculated by the number of hits for a specific video on specific days. First classification is carried out depending on the listener's test and by applying AODE it is again classified which yields higher clarity close to 90%.

The subsequent parameters are supposed for executing with Java programming. Declaration of parameters is embodied in Table 1 and the parameter terms are remarked as Seeds, Peers, Cache, Fragment time, Sendport, Receiveport, Capacity, Packets/ sec. Every parameter holds the smallest and the highest values which can be achieved by basic declaration of Java programming. The threshold worth possesses been requested for every single and every single parameter and larger aftermath attained alongside this value.

TABLE 1: PARAMETERS DECLARATION

Parameter Name	Min	Max	Threshold Value
Seeds	2	50	25
Peers	2	100	50
Cache	1mb	3 mb	3 mb
Fragment time	1 sec	2 sec	2 sec
Sendport	1022	-	-
Receiveport	1900	-	-
Capacity	100	500	300
Packets/sec	1	5	3

TABLE 2: ACCORDING TO LISTENER TEST

Sl.No	Videos Types	Number of Videos	Occur in No. of Server
1	BASICS OF	33	11
2	C++	14	24
3	ORACLE	35	14
4	NS2	27	13
5	VB	27	14
6	MATLAB	14	27
7	VB.NET	6	5
8	LINUX	46	27
9	JAVA	70	25

Table 2 represents mainly three columns. The first column shows video types based on client request which the second column shows the number of video hits made by listeners and the third represents the following videos available in servers.

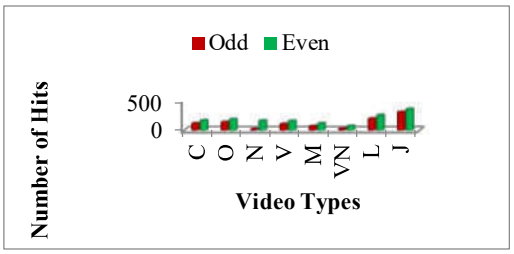


Fig.1: Number of hits for all video in all seasons

Fig.1 has been ready on the premise of the hits generated throughout every season. For example, BASIC OF COMPUTER video gets 3273 hits in odd semester and gets 101 hits in even semester. Embodies an association upheld the server hit for e period and additionally the servers' hits are computed for every single period and drawn in Fig.2.

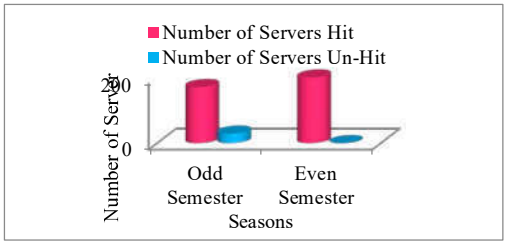


Fig.2: Classification based on the server hit on each season

Table 3 and Fig.3 show the number of hits for each video with the server in both semester. Here is an attempt to classify servers as odd semester and even semester. Based on the semester the result the hit rate will vary. The subjects in the corresponding semester always having solid hit rate. Along with corresponding student hit rate the failed candidate hit rate will combined to give higher hit rate in corresponding semester. The followings are the representations of abbreviations used in Table 3 and Fig.4.

B - BASICS OF COMPUTER; C- C++; O- ORACLE; N- NS2;V- VB
M - MATLAB; VN - VB.NET; L- LINUX; J - JAVA

TABLE 3: NUMBER OF HIT OF EACH VIDEO WITH SERVER IN ALL SEASON

Sl.No	B	C	O	N	V	M	VN	L	J
1	Odd	120	140	15	110	70	25	200	300
2	Even	170	190	166	160	120	75	250	350

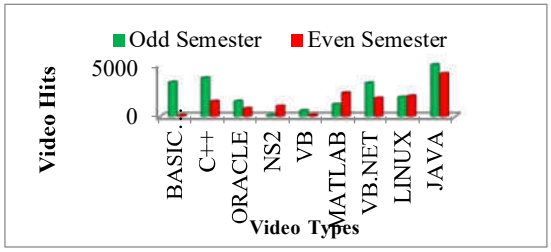


Fig.3: Number of hit of each video with server in all season

V. CONCLUSION

The By using the method of Retrieve, Reuse, Revise and retain, Case-Based Reasoning algorithm categorises the popular videos and separation of video server into seasonal and non-seasonal is performed by Averaged One-Dependence Estimators algorithm. The algorithm shows the JAVA and LINUX domain is non-seasonal with the highest hit or steady hit during all season. Based on this classification, the video service forwards definite server and that particular server offer service to clients. Therefore, by utilizing this kind of algorithms the P2P network provides better service throughout the different seasons and based on the ratios of the season's hits it makes the server available.

REFERENCES

[1]. Narayanan Madeshan, Arun Chokkalingam.: An Efficient Super-Peer Selection for Peer-to-Peer Live Streaming Networks over Video-On Demand Service. J COMPUT THEOR NANOS 13, 7, 4606--4613 (2016).

[2]. Narayanan Madeshan, Arun Chokkalingam. An Efficient Super Peer Selection Algorithm for Peer-To-Peer (P2P) Live Streaming Network. J THEOR APPL INF TECH , 70, 1, 1--8 (2014)

[3]. Narayanan, M., and C. Arun.: Categorize the video server in P2P networks based on seasonal and normal popularity videos using machine learning approach. In Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, pp. 1220--1228. IEEE, Coimbatore(2015)

[4]. Narayanan, Madeshan, and C. Arun. An efficient technique for video content managing in Peer-to-Peer computing using multilevel cache and bandwidth based cluster. In Signal Propagation and Computer Technology (ICSPCT), 2014 International Conference on, pp. 317--322. IEEE, Ajmer(2014)

[5]. Narayanan, M., and C. Arun. An Efficient Method for Handling Data Segment with Multi-Level Caching over Video-on-Demand using P2P Computing. EP J SCI REC, 93, 2 206--213 (2012)

[6]. Narayanan, M., and C. Arun. To Manage Traffic In P2P Networks Using Expectation–Maximization Approaches Over Video-On Demand Services. INT J APPL ENG RES, 10, 4, 2958--2966 (2015)

[7]. Narayanan, M., and C. Arun. Manage Traffic in P2P Live Streaming using Adaptive Routing Scheduling over Video-On Demand Service. ARPN J ENG APPL SCI 10, 13,5581--5587 (2015)

[8]. Narayanan, M., and C. Arun. Outliers Based Caching of Data Segment with Synchronization over Video-on Demand using P2P Computing. RES J APPL SCI ENG TECHNOL 7, 21, 4559--4564 (2014)

[9]. Szkaliczki, T., Eberhard, M., Hellwagner, H, Szobonya, L. Piece selection algorithms for layered video streaming in P2P networks, DISCRETE APPL MATH,167,269--279(2014)

[10]. Lloret, J., Canovas, A., Tomas, J. and Atenas, M. A network management algorithm and protocol for improving QoE in mobile IPTV, COMP COMMUN,35, 15, 1855--1870(2012)

[11]. Kettig, O. and Kolbe, H. J. Monitoring the Impact of P2P Users on a Broadband Operator's Network over Time, IEEE T Network SERV MANAGE, 8,2,116--127(2011)

[12]. Hua, K. L., Chiu, G. M., Pao, H. K. and Cheng, Y. C. An efficient scheduling algorithm for scalable video streaming over P2P networks, COMP NETW, 57, 14, 2856--2868(2013)

[13]. Jeong, M. G., Morrison, J. R. and Suh, H. W. Approximate Life Cycle Assessment via Case-Based Reasoning for Eco-Design, IEEE T AUTOMAT SCI ENG, 12, 2, 716--728(2015)

[14]. Li, H., Zhao, C. and Wu, W. Research on case-based reasoning search method, In The 26th Chinese Control and Decision Conference (2014 CCDC) IEEE, 2360--2364(2014)

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A Neighbor Coverage Protocol with Broadcasting Mechanism for Data Dissemination in Mobile IoT Networks

Puladas Sandhya Priyanka[1] Vasala Tirumala[2] Ediga Lingappa[3]

Assistant Professors, CSE Department,

Malla Reddy College of Engineering, Dhulapally, Kompally, Secunderabad-500100

Abstract - **In recent technology developments, IoT(Internet of Things) has become major role playing in our day-to-day lives. The implementation of IoT applications is the communication between the devices that contain electronics, software, actuators and connectivity that help in the exchanging of data. This paper discusses about the spreading of data among the nodes in the mobile IoT networks. The mechanism is all about having knowledge about the nodes whether they are within the range but have not received any packet and to know whether the packet has to be discarded or to be rebroadcasted to other nodes.**

Key Words: Wireless Sensor Networks, Sensor Nodes, Energy – Efficiency.

I. INTRODUCTION

The world around us is of completely smart way which is of Internet of things (IoT) technology. The communication that was usually among the humans initially but now with the electronic devices, vehicles, computers etc. The intention of these IoT technologies is to sense the data or gather the data from the IoT systems and transmit them to the desired destiny to get processed and analyzed. It is not only to sense the data but to know whether all the communicating nodes are within the range to send the data and to know whether the entire data has been received by the nodes correctly and to know any packet of data has to be rebroadcasted to the nodes. It is important in the communication that knowing all the neighbor nodes to transmit the data. Many mechanisms have been proposed to discover the neighbor nodes in the communication. The figure depicted in Figure 1, shows the various applications where Internet of Things (IoT) have been widely implemented.

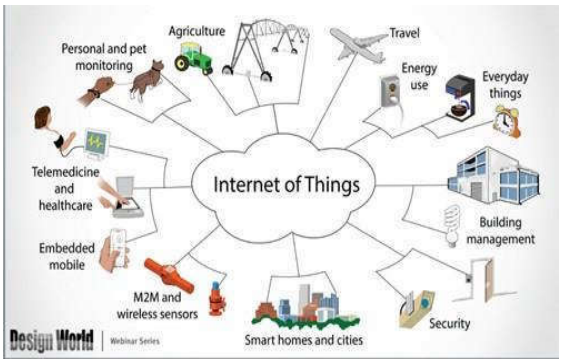


Figure 1: Various Applications of Internet of Things

The following paper discusses about the neighbor discovery and their coverage and also a scheme has been proposed to whether rebroadcast any packet to any node. The rest of the paper is organized as follows: Section II reviews the related works. Section III describes the Proposed mechanisms. Section IV shows the simulation results and the comparisons made. Section V briefly concludes the work done and the future works to be done.

II. RELATED WORKS

In order to send the packets from one node to another node there must be some cooperative mechanisms associated between the nodes and also routing the data from one node to another also shows that whether the nodes are in active state(connected properly) to receive the data. Also some neighbor discovery mechanisms have been implemented to know the nodes are within the network.

2.1 Role of Co-Operative Mechanisms

Need of Co-operative Mechanisms: In order to have proper transportation of data among the nodes

ISBN 978-81-939386-1-4

in the network, there must be cooperation among the nodes. Cooperation mechanism leads to establish sending of data where no data is lost. The following lines make us clear in knowing how the cooperative mechanism is carried out.

Various Co-operative Mechanisms: The mechanism that is carried out by the coalition-based data transport mechanism where neighboring nodes are organized into groups to form coalitions and sensor nodes within one coalition carry out cooperative communications [6]. The other scheme is that the nodes choose the efficient path selection by relay nodes and then transport the data to the next node [10]. If the medium is of broadcast nature then the network adapts two phase model where in one phase the data transport is sent to the receiver node and if the data is not received by that node then in second phase it is sent through the relay node [9]. Also the cooperation can be based on the two protocols called DRP (Data Reservation Protocol) and PCA (Prioritized Channel Access) where the slot reservation take place and hence the collisions are avoided [4].

2.2 Role of Node Discovery Mechanisms

Need of Node Discovery Mechanisms: One of the important design issues of the mobile networks is to have knowledge about the neighbor nodes and the proper connectivity among them in the network. The following lines describe how the node discovery mechanism is being implemented so far.

Various Node Discovery Mechanisms: By sending the beacon messages the nodes present in the network can be known [3]. Also target discovery plays a major role as the node need to know about the nearby node to which it has to send the data. Thereby, the nodes can be classified into disjoint sets so as to perform the tasks in an efficient way [2]. By adapting the selective communication mechanism the node can know about the efficient path in which nodes are capable of handling the data when routed and by having prioritization of the data high priority data can be sent fast whereas low priority can be discarded [8].

2.3 Role of Routing Mechanisms

Need of Routing Mechanisms: To send the data from one node to another, series of instructions to be followed, so that there exists some path between the nodes to transfer the data to reach the destination. Many routing mechanisms have been proposed by many researchers. The following lines describe about the various mechanisms implemented so far.

Various Routing Mechanisms: In routing the packet, the location of a node to which the sender node has to be sent must be known and thereby, location aware comes into existence since all the nodes in the network will be in mobile in nature [1]. When the network is considered with the cluster formation for performing the tasks in more efficient manner, Table maintenance and the cluster head maintenance comes into the picture in order to maintain the consistency of the nodes and their energy information [5]. A New Multipath Routing Approach is also proposed in which Route Request control messages can be sent by and controlled [7].

2.3 Broadcasting Mechanisms

A gossip based broadcast mechanism was proposed that describes each node rebroadcasts a newly received packet [11]. A Counter based strategy tells that one particular node named as counter, will be keeping track of the same data packet received by the node. The value is then compared with the initial value and the threshold value reaches, the rebroadcasting value is made to zero else one[12]-[14]. Location based strategy was also implemented in [15] and [16] based on the GPS locations of the nodes in the network, the data packets are rebroadcasted.

III. NEIGHBOR COVERAGE PROTOCOL WITH BROADCASTING MECHANISM FOR DATA DISSEMINATION

a) Neighbor Coverage: It is necessary that in all the networks where the data has to be sent from one node to another node there must be proper connectivity among the nodes. Hence this can be achieved by one mechanism called Neighbor Coverage.

This mechanism is all about nodes in the networks give their presence in the network with an initial packet named as HELLO Packet so that each node will be knowing each other node's presence and thereby, finding out the rebroadcasting delay of a packet to be sent.Also, that there must one data

packet which has to carry the neighbor list of the transmitter.

b) Broadcasting Mechanism:

During the transfer of data packets from one node to another each node will be maintaining the neighbor list of the transmitter and also the nodes which received the packet. By doing so, two things happen one is rebroadcast delay can be calculated and also data packet availability with the node. Thereby, the packet delivery ratio can be calculated from the availability of a particular data packet.

c) Adaptability:

Each time when a node enters into the packet should send a HELLO packet and thereby, the adaptive nature of the network would be as the ratio of newly added nodes to the already existing nodes in the network. This can be configured by the system administrator.

Rebroadcast delay = time allotted for the packet to rebroadcast+ additional time needed

Additional time = Number of nodes already in the network/ Newly Added nodes.

Algorithm:

IV. CONCLUSION

As the technology has grown into the automatic mechanism, the sensors are part of it. Thereby, several applications of the sensors have been taking place in various fields like – hospitals, crime investigation by vehicle tracking, military, civil engineering, etc. In order to increase the application of sensors widely, the major thing is to design the sensor networks in a more efficient and effective way. In this regard, the major problem faced by the sensor networks is the energy degradation of the nodes which are battery made up and have limited amount of energy which is also not be able to recharged further. All these mechanisms indicate how to maintain the efficiency and the performance of the network by reducing the energy consumption of the nodes eventually of the entire network. These help the researchers to perform research and increase the performance levels of the wireless sensor networks which would be the most benefitted part in the recent technology development where sensor world would be seen in the very few years of the present generation.

REFERENCES

[1] C.Shanti and D.Sharrmila "A self-organized location aware energy efficient protocol for wireless sensor networks" in Elsevier Computers and Electrical Engineering,2014.

[2] Mihaela Cardei, My T Thai Yingshu Li and Weili Wu "Energy-Effcicient Target Coverage in Wireless Sensor Networks" in IEEE (0-7803-8968-9), 2005.

[3] Mikko Kohvakka, Jukka Suhonen, Mauri Kuorilehto, Aille Kaseva, Marko HAnnikainene and Timo D.Hamalainen " Energy Efficient neighbor Discovery prototcol for mobile sensor networks", in Elsevier Ad hoc Networks, 2009.

[4] M.Rengasamy, E.Dutkiewicz and M.Hedley "MAC Design and Analysis for Wireless Sensor Networks" in International Symposium on Communication Information Technologies, 2007.

[5] Mudasser Iqbal, Iqbal Gondal and Laurence Dooley "A cross-layer Dissemination protocol for Energy Efficient Sink Discovery in wireless sensor networks" in IEEE for publication in ICC proceedings 2007.

[6] Qinghal Gao, Junshan Zang, Xuemin and Bryan Larish "A Cross-Layer optimization Approach for Energy-Efficient Wireless Sensor Networks: Coalition-Aided Data Aggregation, Co-operative Communication, and Energy Balancing" in Hindwai Publication, 2007.

[7] Saira Banu and R.Dhanasekaran "A New Multipath Routing Approach for Energy Efficiency in Wireless Sensor Networks" in International Journal of Computing Applications (0975-8887), 2012.

[8] Sara Pino-Povedano, Rocio Arroyo-Valles and Jesus Cid-Sueiro "Selective Forwarding for energy-efficient target tracking in Sensor Networks" in Elsevier Signal Processing, 2014.

[9] Shaoqing Wang and Jingnam Nie " Energy Efficiency Optimization of Co-operative Communication in Wireless Sensor Networks", in Hindwai Publication 2010.

[10] Weiwei Fang, Feng Liu, Fangnan Yang, Lei Shu and Shojiro Nishio, "Energy – Efficient Cooperative Communication for Data Transmission in Wireless Sensor Networks", in IEEE 2010, (0098 3063), Page No: 2185-2192.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

CONTENT BASED CONFIDENTIALITY DETECTION METHOD FOR DATA LEAKAGE PREVENTION

Subhashini Peneti
Assistant Professor
Department of CSE
Malla Reddy College of Engineering
Hyderabad, India
subhashinivalluru@gmail.com

Padmaja Rani B
Professor
Department of CSE
Jawaharlal Nehru Technological Universty
Hyderabad,India.
Padmaja_jntuh@jntuh.ac.in

Abstract— **Protecting confidential data became a challenge for all private and public organizations. According to Gartner report, the majority of data leakages in organizations are due to internal factors. Data Leakage Prevention Systems can protect monitor and identify the confidential data at-rest, in-use and in-motion. This paper presents Data Leakage Prevention system, to prevent confidential data from leakages using the Term Based Confidentiality Detection Method .The proposed method consists of two phases: training and testing phase. The training phase identifies confidential terms from the documents and testing phase detects the confidentiality of the document.**

Keywords— confidential, data, data leakages, security, organizations.

I. INTRODUCTION

Now a day's modern enterprises depends on data sharing in both inside and outside the organization. Data sharing increased the data leakages [1]. The data which causes data leakages are SSN (social security numbers), medical records, organization financial information and trade secrets. In order to prevent confidential data from the leakages, organizations should focus of their internal hosts and network parameters. Organization security measures can also prevent confidential data from classic threats like viruses, Trojan horses, worms, D/Dos attack and intrusions. In addition organizations are required to meet the terms and regulations of the state and central government.

Data leakage is defined as the accidental or unintentional distribution confidential data to an unauthorized entity [2].In information security, identification of data leakage attack has become a critical issue for every organization [2]. Most of the data leakage attacks are caused from employees of the organization.

Organization must keep a close monitor to secure their confidential data. Data Leakage Prevention (DLP) solution is one of the recent methodology and technical solution for confidential data security .Data leakage prevention solution provides security to the confidential data from the inside and outside the network.

Content-aware DLP solution is a data leakage prevention solution that involves awareness of the content that is being protected. The content-aware DLP solution protects organizations from data leakage threats.

Content-aware DLP solutions are able to read the entire content of the document and identify confidential data, including the text found in the organization's documents. Content-aware DLP solutions detect and prevent the confidential data which is available in motion, in use and at rest [13].

II. RELATED WORK

Different types of DLP solutions are available in the market, some of the solutions protecting confidential data by providing tag to the data i.e. the organization documents are classified in to confidential and non-confidential with tags(confidential and non-confidential), if any employee in the organization sends a new document then identify the tag of the document. The document is blocked, when the document contains a confidential tag otherwise forwarded to the network.

DLP solutions are categorized based on the state of the data. In a DLP the data is available in three states [3]: DAR(Data-At-Rest), DIU(Data-In-Use) and DIM(Data-In-Motion). DAR is defined as the whole amount of data available in the organization's data centers. The following table represents the summary of the DLP solutions for Data-At-Rest.

TABLE 1. SUMMARY OF THE DLP SOLUTIONS FOR DATA-AT-REST

Category	Method	Proposed by, year	Issues	References
Misuse detection in Information retrieval system	User profiling based on query results and relevance feedback	Rebecca Cathey, Nazil Goharian and David Grossman ,2003	Requires huge amount of profiling	[4]
	User profiling based on keyword in queries and search	Ling Ma and Nazil Goharian, 2005	Requires regular administrative referen	[5]

	results(relevance feedback)		ces	
	Assigning privilege levels to users and security level to documents and monitoring user access to documents.	Hyeran Mun, Kyusuk Han and Kwangjo Kim,2014	If a leak is happening the method is ineffective	[6]
Misuse detection in Database	Using machine learning methods to detect abnormal access behavior by analyzing query syntax.	Ashish Kamra, Evimaria Terzi, and Elisa Bertino,2008	Lacking query semantics.	[7]
	Using machine learning methods to detect abnormal access behavior by analyzing query result-set.	Sunu Mathew , Michalis Petropoulos , Hung Ngo , Shambhu Upadhyaya,2009	Proposed method work for static database i.e. no updates	[8]
Data hidden in files	Awareness	Simon Byers ,2004		[9]
Encryption and access control	Encryption and access control.	Imad M. Abbadi, Muntaha Alawneh, 2008	Encryption can secure confidential data	[10]
	Application of selective encryption to provide selective access control to outsourced confidential data by third-party partners.	Sabrina De Capitani ,Di Vimercati , Sara Foresti,2010	Does not detect data leak	[11]

DIU is the data that user is interacting. The following table represents the DLP solutions for DIU.

TABLE 2. SUMMARY OF THE DLP SOLUTIONS FOR DATA-IN-USE

Category	Method	Proposed by, Year	Issues	References
Detecting Malicious Insiders using Honey pots	Web-based service generated and distributed decoy documents to registered users and monitors by	Brian M. Bowen, Shalom Hershkop, Angelos D. Keromytis and Salvatore J. Stolfo, 2009	Produces overheads	[12]
	Insertion of a honey table into database which is able to attract malicious users.	Antanas Čenys, Darius Rainys, Lukas Radvilavičius, and Nikolaj Goranin,2005	Honey table Creation	[13]
	Distributes confidential data objects to several agents	Maya Bercovitch, Meir Renford, Lior Hasson Asaf ,Lior Rokach and Yuval Elovici,2011	Limited to specific situations or scenarios.	[14]

The data that is being transmitted through the network is called as Data in Motion (DIM). The following table presents the DLP solutions for DIM.

TABLE 3.SUMMARY OF DLP SOLUTIONS FOR PROTECTING DATA-IN-MOTION

Category	Method	Proposed by, year	Issues	References
	Different rules are framed , these rules determine the confidentiality of the document	W.W. Cohen,1996	Produce high False Positive Rate	[15]
	Automatic text classification algorithms for classifying enterprise documents as either confidential or non-confidential	Michael Hart, Party's Manadhata and Rob Johnson ,2011	Not able to classify encrypted and multimedia content	[16]

	Analysis of the email exchange among groups of members of the organization to detect email leakage incidents	Polina Zilberman, Asaf Shabtai and Lior Rokach, 2011	False positive leaks because of a short history between a sender and a recipient	[17]
	Context based confidentiality detection method(CoBAn) using machine	Gilad Katz, Yuval Elovici and	Produce high False Positive Rate	[18]
Network/Web-based protection	Monitoring user's access to information on an intranet using network-based sensors which generates information-use events. These events are combined with contextual information and processed by various rule-based and statistical detectors that may issue alert.	Caputo, 2009	Not able to classify encrypted and multimedia content	[19]

It is observed that, the above all methods which are proposed for email leakage protection are suffered from a high false positive rate. In contrast to the above methods, we proposed a method to prevent confidential data from the leakages with low false positive. The proposed method "Term Based Confidentiality Detection Method" protects data in-motion and used content-based inspection for preventing data leakages. CoBAn (**Co**ntext **B**ased **An**alysis) method is taken up as a baseline method[18].

III. TERM BASED CONFIDENTIALITY DETECTION METHOD

Term Based Confidentiality Detection Method is a content-based confidentiality detection method which identifies the confidentiality of the test document based on the content of the document. This method classifies organizational documents into two classes: Confidential and Non-confidential by using the content of the document.

TBCDM consists of two phases: training and detection phase. During the training phase, clusters of documents with confidential terms are generated, in the detection phase, each tested document is assigned to clusters and its content are then matched to each cluster's respective confidential terms to determine the confidentiality of the document.

The proposed method maintains the confidential and non-confidential documents separately and identifies the confidential data through confidential terms. Confidential terms serve as an initial indication of the presence of confidential content in the document. The proposed method uses language modeling technique for confidential terms identification.

Term based confidentiality detection method consists of two phases

1) Training phase and
2) Testing phase.

A. Training Phase

The objective of the training phase is to represent the confidential content of documents using the training repository. The training repository consists of both confidential and non-confidential documents of the organization.

The main operation in the training phase is language model creation. The language model is created separately for confidential clusters set and non-confidential clusters set. Language model consist of the following steps:

1. Document pre-processing
2. Unsupervised cluster creation
3. Confidential terms detection

a) Document Pre-processing

In general the document repository is not available in understandable format; with the help of data pre-processing the document repository is converted into understandable format [21].

b) Unsupervised Cluster Creation

To find the confidentiality of a document, we need to identify the various subjects represented by organization documents. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled. The most common unsupervised learning method is cluster analysis [22].

For cluster Creation, the content of the every document is represented as a vector using TF-IDF (Term Frequency-Inverse Document Frequency).

In information retrieval, TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection[20].

TF-IDF (term) = TF (term) *IDF (term).

Where TF is the term frequency of a term in a document and IDF is the inverse document frequency of a term.

After TF-IDF calculation of each document in the training repository, the next step is to generate confidential and non-confidential clusters. For cluster creation K-means unsupervised algorithm with cosine measure as the distance function is used.

c) Confidential Terms Detection

The objective of the confidential terms detection method is to identify the terms, which indicate with a high probability in a confidential cluster and low probability in a

non-confidential cluster, we referred these terms as confidential terms [23].

The Confidential terms have two purposes:

- They serve as initial indicators of relevant content; and
- They help to find context terms.

Our first aim is to find the terms with a high probability of appearing in confidential documents and a low probability of appearing in non-confidential documents. In order to find the confidential terms, performs two operations:

1. Probability calculation of each term in confidential and non-confidential clusters (CR, NR).
2. Confidential score calculation for each term of CR

1) Probability Calculation

The probability of a term indicates the occurrences of a term in the cluster [83]. The higher the probability of a term indicates the more occurrence of the term in the cluster. Calculate the probability of each term available in the confidential and non-confidential clusters using the following formula,

$$p_{lm}(t/c)=(t_f)/N(t,c) \quad (1)$$

Where $p_{lm}(t/c)$ is the probability of a term t in cluster C,tf_t is a term frequency of a term t in a cluster C and N (t, c) is the total number of terms in C. In a cluster if a term appears more than one time then add all the probabilities of a term and maintain only one probability .Once the probability calculated for all terms of confidential and non-confidential clusters, calculate confidential score for each term of CR.

2) Confidential Score Calculation

To calculate the confidential score of a term we need two inputs: Confidential clusters with term probabilities (CR_{LM}) and Non-confidential clusters with term probabilities (NR_{LM}). In order to find the occurrence of a term in non-confidential and evaluating its effect towards confidentiality can be measured by finding similar Non-Confidential clusters to a corresponding confidential cluster. The term occurrence and its weight age among similar non-confidential documents can be measured in one of the following three ways. Three methodologies are proposed for finding the confidentiality of a term.

1. In method I, the combined effect of a term occurrence in non- confidential clusters compared to term occurrence in a confidential cluster is considered. Hence, the sum of all probability of occurrence of a term in non-confidential is considered because if the term appears in more than one non- confidential cluster its combine effect on confidentiality is less.
2. In method II, the average effect of a term occurrence in all non- confidential clusters compared to the term occurrence in a confidential cluster. For this reason the average of all probability of occurrence of a term in non-confidential is considered.
3. In method III, the effect of a term occurs maximum in non- confidential clusters that are similar to confidential cluster is considered. Hence the maximum occurrence of a term out of all probability occurrence of that term in non-confidential cluster compared to confidential cluster is considered.

Method I: Confidentiality of a term is measured in terms of term probability of occurrence in confidential documents divided by the summation of term probability of occurrences in all similar non-confidential clusters.

$$\left[confidential\_score(t) = \frac{p\left(t/_{cr}\right)}{p\left(t/_{nr1}\right)+p\left(t/_{nr2}\right)+p\left(t/_{nr3}\right)+\ldots\ldots+p\left(t/_{nrn}\right)} \right] \quad (2)$$

Where Cr is the analyzed confidential cluster, is the probability of a term t in Cr, nr is the non-confidential cluster that is similar to Cr and $p\left(t/_{nr}\right)$ is the probability of a term t in nr.

$$\text{Let } p\left(t/_{cr}\right) = p_{cr} \quad (3)$$

$$p\left(t/_{nr1}\right)= p_{nr1},\ p\left(t/_{nr2}\right)= p_{nr2},\ p\left(t/_{nr3}\right)= p_{nr3},\ldots\ldots\ldots\ p\left(t/_{nrn}\right)= p_{nrn} \quad (4)$$

Using eq. (3) and eq. (4)

$$confidential\_score(t) = \frac{p_{cr}}{p_{n_{r1}}+p_{n_{r2}}+p_{n_{r3}}+\ldots\ldots+p_{n_{r3}}} \quad (5)$$

$$confidential\_score(t) = \frac{p}{\sum_{i=1}^{n} p_{nri}} \quad (6)$$

Method II: The confidentiality of a term is measured in terms of term probability of occurrence in confidential documents divided by an average of term probability of occurrence in all non-confidential clusters that are similar to a corresponding confidential cluster.

$$confidential\_score(t) = \frac{p\left(t/_{cr}\right)}{\frac{p\left(t/_{nr1}\right)+p\left(t/_{nr2}\right)+p\left(t/_{nr3}\right)+\ldots\ldots+p\left(t/_{nrn}\right)}{N}} \quad (7)$$

Where Cr is the analyzed confidential cluster, $p\left(t/_{cr}\right)$ is the probability of a term t in Cr, nr is a non-confidential cluster that is similar to Cr, $p\left(t/_{nr}\right)$ is the probability of a term t in nr and N is the number non-confidential clusters that contain term t.

Using eq. (3) and eq.(4)

$$confidential\_score(t) = \frac{p_{cr}}{\frac{p_{nr1} + p_{nr2} + p_{nr3} + \ldots\ldots + p_{nrn}}{N}} \quad (8)$$

$$confidential\_score(t) = \frac{N * p_{cr}}{p_{nr1} + p_{nr2} + p_{nr3} + \ldots\ldots + p_{nrn}} \quad (9)$$

Method III: Confidentiality of a term is measured in terms of term probability occurrence in confidential documents divided by maximum value of term probability occurrence of a term t in all similar non-confidential clusters.

$$confidential\_score(t) = \frac{p\left(t/cr\right)}{Max\left[p\left(t/nr1\right), p\left(t/nr2\right), p\left(t/nr3\right), \ldots\ldots, p\left(t/nrn\right)\right]} \quad (10)$$

Where Cr is the analyzed confidential cluster, $p\left(t/cr\right)$ is the probability of a term t in Cr, nr is the non-confidential cluster similar to Cr and $p\left(t/nr\right)$ is the probability of a term t in nr.

Using eq. (3) and eq. (4)

$$confidential\_score(t) = \frac{p_{cr}}{Max\left[p_{nr1}, p_{nr2}, p_{nr3}, \ldots\ldots, p_{nrn}\right]} \quad (11)$$

Comparison of three methods:

Considering Method I and II,

In Method I,

$$confidential\_score(t) = \frac{p_{cr}}{\sum_{i=1}^{n} p_{nri}}$$

In Method II,

$$confidential\_score(t) = \frac{p_{cr}}{\frac{p_{nr1} + p_{nr2} + p_{nr3} + \ldots\ldots + p_{nrn}}{N}}$$

Since

$$\sum_{i=1}^{n} p_{nri} > \frac{p_{nr1} + p_{nr2} + p_{nr3} + \ldots\ldots + p_{nrn}}{N} \quad (12)$$

$$\frac{p}{\sum_{i=1}^{n} p_{nri}} < \frac{N * p_{cr}}{p_{nr1} + p_{nr2} + p_{nr3} + \ldots\ldots + p_{nrn}} \quad (13)$$

Considering Method I and III

In Method I,

$$confidential\_score(t) = \frac{p_{cr}}{\sum_{i=1}^{n} p_{nri}}$$

In method III,

$$confidential\_score(t) = \frac{p_{cr}}{Max\left[p_{nr1}, p_{nr2}, p_{nr3}, \ldots\ldots, p_{nrn}\right]}$$

Let

$$\left[Max\left[p_{nr1}, p_{nr2}, p_{nr3}, \ldots\ldots p_{,nrn}\right] = p_j\right] \quad (14)$$

Since

$$p_{nr1} + p_{nr2} + p_{nrj-1} + p_{nrj} + p_{nrj+1} + \ldots\ldots + p_{nrn} > p_j$$

$$\frac{p_{cr}}{\sum_{i=1}^{n} p_{nri}} < \frac{p_{cr}}{p_j} \quad (15)$$

From the above observations(eq.(13) and eq. (15),When method 1 is compared with method II and method III , the method II and method III results in high confidential score hence results in more FPR than method I. So, for further research evaluation Method I is chosen as a proposed method for confidential score calculation.

B. Testing Phase

The main objective of the detection phase is to detect the confidentiality level of the test document in terms of a score [84]. In order to find the confidential score of the tested document, apply data pre-processing steps, transformed into TF vector format, find the similar confidential clusters and calculate the confidentiality value of the test document.

Testing phase consists of the following steps:

1. *Assign the test document to a relevant confidential cluster:* The purpose of this step is to identify which of the confidential clusters to be used to determine the confidentiality level of the tested document. This is done by using the cosine similarity measure (as discussed in section 3.3.2.3); all the confidential clusters whose similarity is greater than a predefined threshold are selected.
2. *For each of the similar cluster, identify all the confidential terms that appear both in the tested document and clusters.*

Scan the text of the test document and identify the terms that appeared both in test documents and similar clusters. Consider the terms of similar clusters, whose score is greater than 1 to be included in the document's confidentiality calculation.

3. *Calculate the document's confidentiality score.*

For each similar cluster, calculate the score by summing up the scores of all the confidential terms. If

ISBN 978-81-939386-1-4

more than one similar clusters available then the confidentiality of the tested documents is calculated by a summation of scores of all the similar clusters.

Confidentiality value (similar cluster) = summation of scores of all confidential terms that appear both in the tested document and similar cluster.

Confidentiality value (tested document) = summation of all similar clusters score.

4. *Determine whether the document is confidential.*

If the confidentiality value of the tested document is above the threshold then that document marked as a confidential and it is blocked.

C. Algorithms

This section presents different algorithms used for the proposed method. Training phase and Detection phase algorithms are the two main algorithms in the proposed method.

a) Training Phase Algorithm

The training phase algorithm is to create confidential clusters with confidential terms by taking confidential, non-confidential documents as an input.

Training phase algorithm

Input

1. List of confidential documents
2. List of non-confidential documents

Output

CTR – List of confidential clusters, each cluster with confidential terms and their scores.

Steps:

1: CD ← Read set of confidential documents

2: ND ← Read set of non-confidential documents
3: T ← Read confidential score threshold value
4: Perform clustering on confidential documents with K-means unsupervised clustering algorithm and assign set of confidential clusters to CR
// CR is a list of confidential clusters
5: Perform clustering on non-confidential documents with K-means unsupervised clustering algorithm and assign set of confidential clusters to NR
//NR is a collection of K non-confidential clusters
6: For each cluster cr CR
Calculate probability of each term and write into CR_{LM}
End for
// CR_{LM} is a term probability values of all terms in a cluster CR
7: For each cluster n_r NR
Calculate probability of each term and write into NR_{LM}
End For
//$NR_{LM}$is a term probability values of all terms in a cluster n_r
8: For each cluster cr CR_{LM}
For each term t in cr
Calculate confidential score
If(score (t) >T)
Consider the term t as a confidential term and assign term t and its score to CTR
End if
// CT is a set of confidential terms and their scores of a cluster CR
End For
End for

b) Confidential Score Calculation of a Term

This algorithm is to identify the terms which are having high probability in confidential documents and low confidentiality in the non-confidential documents.

Input:

1. List of confidential clusters with probability values.
2. List of non-confidential clusters with probability values.
Score calculation threshold

Output:

CTR - Set of clusters,each with confidential terms and their scores.

Steps:

1:CR_{LM}← read set of confidential clusters with probability
2: NR_{LM}← read set of non-confidential clusters with probability
3: TR – initialize similarity threshold
4: For each cluster Cr in CR_{LM}
NSR_{LM}←Find Non-confidential clusters whose cosine similarity to CR is above TR
Term tis available in a more than one similar cluster, add probabilities and maintain one probability for a term.
For each term t in Cr
Calculate Confidential Score (t) //t available in C

// term not available in non-confidential clusters
Confidential Score (t) = 0.01
If score greater than 1
CTR [CR] ←CR[t]
CTR [CR]← score [t]
// CTR is a set of confidential cluster, each with confidential terms and their scores
End if
End for
Return t
End for

c) Testing Phase Algorithm

The testing phase algorithm is to find the confidentiality of the test document in terms of score. If score greater than a predefined threshold then consider that test document is confidential otherwise non-confidential.

Algorithm

Input:

1. The test document whose confidentiality we wish to determine
2. Set of confidential term clusters

3 Confidentiality threshold to determine the confidentiality of the tested document
4.Similarity threshold, to detect the similar clusters
Output:
Conf-score – represents the confidentiality (yes/no) of the tested document
Steps:
1: Read test document D
2: Read confidential clusters, CTR
3: Read confidential score threshold, T_C
4: Read similarity threshold, Tr
5: InitializeConf_score = 0
6: For each cluster CR CTR
Cosvalue ← similarity between D and CR
If (cosvalue>Ts)
add CR to p as a similar cluster// p is a set of similar clusters
End if
End for
7: For each cluster CR∈p
Initialize score = 0 // score of a cluster CR
Find all terms that appear both in D and CR and assign to M
For each term t∈ M
Score = score + score (t)
End for
Conf_Score= Conf_score+score
End for
8: If (Conf_score> T_c)
Test document marked as a confidentinal
Else
Test document marked as a Non-confidential
End if

D. Evaluation of TBCDM

This section presents the evaluation of the Term Based Confidentiality Detection Method (TBCDM) using Reuters and Enron email dataset.

A. Evaluation of TBCDM using Reuters News Article Dataset

Experimental results of the proposed method using Reuters news article dataset are presented in this section[24].

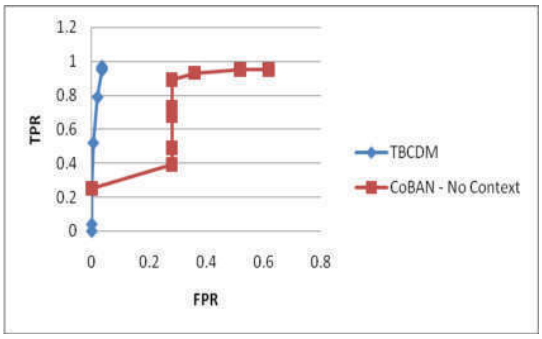


Fig 1.Comparison of TPR and FPR between TBCDM and CoBAn-No context using Reuters News Article.

From the above graph, the performance of the TBCDM is better than the performance of the baseline method CoBAn-No Context. Proposed method performed well for confidentiality threshold 0.05 and similarity threshold 100.

B. Evaluation of TBCDM using Enron Email Dataset

The following is the experimental results of the proposed method using Enron email dataset[25].

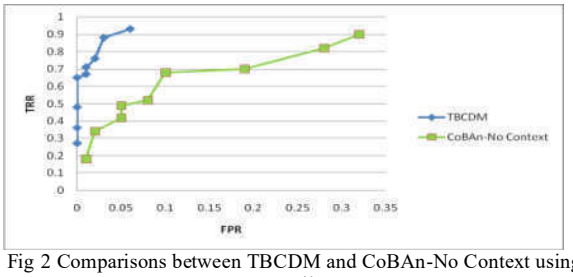


Fig 2 Comparisons between TBCDM and CoBAn-No Context using Enron Email Dataset

For 0.05 confidentiality threshold and 100 similarity threshold, the proposed method TBCDM and baseline method CoBAn-No Context performed well. The above graph shows, the performance of the TBCDM better than the baseline method for all threshold values.

IV. Conclusion.

The present paper described a new confidentiality detection method called "TBCDM" for data leakage prevention. TBCDM detects the confidential content of the document and classifies test documents in to confidential and non-confidential. This paper presents the detailed methodology of the TBCDM .The performance of the proposed method has been measured with different similarity threshold values and different confidential score threshold values.TBCDM detects the confidentiality of the test documents with high true positive rate and low false positive rate. The performance of the proposed method is compared with the base line method (**CoBAn-No Context)**, by using Reuter's news article and Enron email data sets.

REFERENCES

[1]. Michael Hart , Pratyusa Manadhata and Rob Johnson, "Text Classification for Data Loss Prevention", Lecturer notes in computer science ,11th International symposium on pririvacy enhancing technologies ,PETS,2011, PP 18-27.

[2]. Yuri Shapira, Bracha Shapira and Asaf Shabtai, "Content based data leakage detection usingextended fingerprinting, 2006.

[3]. Shabtai, A., Elovici, Y., Rokach, L," A Survey Of Data Leakage Detection And Prevention Solutions", Springer Briefs in Computer Science, Springer, 2012.

[4]. Cathey Rebacca, Ma L and Goharian N, "Misuse Detection for Information Retrieval System". Proceedings, 12th ACM conference on Information and Knowledge Management, 2003.

[5]. Ma L and Goharian N, "Query Length Impact On Misuse Detection In Information Retrieval Systems", proceedings, ACM symposium on Applied Computing, pp 1070-1075.

[6]. Mun H, Han K, Yeun C and Kim K, "Yet another Intrusion Detection System against Insider Attacks", symposium on Cryptography and Information security.

[7]. Karma, a Terzi, Evimaria and Bertino E,"Detecting Anomalous Access Patterns in Relational Databases", International Journal on Very Large Databases, 2008, 17(5), pp 1063-1077.
[8]. Sunu, M Michalis P, Hung N and ShambhuU,"A Data Centric Approach to Insider Attack Detection in Database Systems." Technical Report,2009.
[9]. Simon D Byers "Information Leakage Caused By Hidden Data In Published Documents" IEEE Security And Privacy, Volume 2, Issue 2,March 2004, Pp 23-27
[10]. Abbadi, I.M., and Alawneh, M," Preventing Insider Information Leakage For Enterprises", Proceedings, International Conference on Emerging Security Information, Systems and Technologies, pp 99–106, 2008.
[11]. De Capitani , Vimercati S, Foresti S, Jajodia S, Paraboschi S, and Samarati, P," Encryption Policies For Regulating Access To Outsourced Data", ACM Transactions on Database Systems, 35(2), 12:2–12:46, 2010.
[12]. Bowen B.M, Hershkop S, Keromytis A.D, and Stolfo S.J," Baiting Inside Attackers Using Decoy Documents," Proceedings, 5th International ICST Conference (SecureComm'09,), 2009.
[13]. Čenys, Rainys D., Radvilavičius L, and Gorani, N," Implementation of Honeytoken Module in DBMS Oracle 9ir2 Enterprise Edition for Internal Malicious Activity Detection", 2005.
[14]. Maya Bercovitch, Meir Renford, Lior Hasson Asaf "HoneyGen: An automated honeytokens generator", : 2011 IEEE International Conference on Intelligence and Security Informatics, ISI 2011, Beijing, China, 10-12 July, 2011
[15]. Cohen W.W, "Learning rules that classify e-mail", Proceedings, AAAI Symposium on Machine Learning in Information Access, pp 18–25, 1996.
[16]. Michael Hart , Pratyusa Manadhata and Rob Johnson, "Text Classification for Data Loss Prevention", Lecturer notes in computer science ,11th International symposium on pririvacy enhancing technologies ,PETS,2011, PP 18-27.
[17]. Zilberman, Polina, Shabtai, Asaf, Rokach Lior "Analyzing Group Communication for Preventing Data Leakage via Email", .IEEE, 2011.
[18]. Katz Gilad, Elovici Yuval, Shapira Bracha, "CoBAn: A Context Based Model for Data Leakage Prevention". Elsevier, Journal of Information Science, 262(2014), pp 107-128.
[19]. Caputo D.D, Stephens G.D, and Maloof M.A,"Detecting insider theft of trade secrets", IEEE Security and Privacy, 7(6), pp 14–21, 2009.
[20]. Salton G, Buckley C, "Term-weighting approaches in automatic text retrieval", Information Processing and Management, Vol. 24, No. 5, pp 513-523, 1988.
[21]. A Anguera, Jon Barreiro and J. A Lara,"Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry", Computational and structural Biotechnological journal, Elsevier, Vol 14, pp 185-199, 2016.
[22]. Steinbach M, Karypis G, and Vipin Kumar," A Comparison of Document Clustering Techniques", A Comparison of Document Clustering Techniques. Technical Report #00-034.
[23]. Gregory Glen Odom and La Grand Prairie, "Method and Apparatus for Secured Transmission of Confidential Data Over an Unsecured Network", United States Patent, Oct. 30, 1997.
[24]. **https://archive.ics.uci.edu/ml/datasets/reuters_-** 21578 + text + categorization + collection.
[25]. http://www.cs.cmu.edu/enron

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

Identity-Based Encryption with Cloud Revocation Authority and Its Applications

P.Pravalika Reddy,
M.Tech.Scholar,
Dept. of Computer Science & Engineering,
Malla Reddy College of Engineering

Prof.Kannagi,
Professor,
Dept. of Computer Science & Engineering,
Malla Reddy College of Engineering

Abstract—Identity-based encryption (IBE) is a public key cryptosystem and eliminates the demands of public key infrastructure (PKI) and certificate administration in conventional public key settings. Due to the absence of PKI, the revocation problem is a critical issue in IBE settings. Several revocable IBE schemes have been proposed regarding this issue. Quite recently, by embedding an outsourcing computation technique into IBE, Li *et al.* proposed a revocable IBE scheme with a key-update cloud service provider (KU-CSP).
However, their scheme has two shortcomings. One is that the computation and communication costs are higher than previous
revocable IBE schemes. The other shortcoming is lack of scalability in the sense that the KU-CSP must keep a secret value for each user. In the article, we propose a new revocable IBE scheme with a cloud revocation authority (CRA) to solve the two shortcomings, namely, the performance is significantly improved and the CRA holds only a system secret for all the users. For security analysis, we demonstrate that the proposed scheme is semantically secure under the decisional bilinear Diffie-Hellman (DBDH) assumption. Finally, we extend the proposed revocable IBE scheme to present a CRA-aided authentication scheme with period-limited privileges for managing a large number of various cloud services.

Index Terms—Encryption, authentication, cloud computing, outsourcing computation, revocation authority.

1 Introduction

Identity (ID)-based public key system [1], [2] is an attractive alternative for public key cryptography. ID-PKS setting eliminates the demands of public key infrastructure (PKI) and certificate administration in conventional public key settings. An ID-PKS setting consists of users and a trusted third party (i.e. private key generator, PKG). The PKG is responsible to generate each user's private key by using the associated ID information (e.g. e-mail address, name or social security number). Therefore, no certificate and PKI are required in the associated cryptographic mechanisms under ID-PKS settings. In such a case, ID-based encryption (IBE) allows a sender to encrypt message directly by using a receiver's ID without checking the validation of public key certificate. Accordingly, the receiver uses the private key associated with her/his ID to decrypt such ciphertext. Since a public key setting has to provide a user revocation mechanism, the research issue on how to revoke misbehaving/compromised users in an ID-PKS setting is naturally raised.

Related Work

In 2001, Boneh and Franklin [2] proposed the first practical IBE scheme from the Weil pairing and suggested a simple revocation method in which each non-revoked user receives a new private key generated by the PKG periodically. A pe- riod can be set as a day, a week, a month, etc. A sender uses a designated receiver's ID and current period to encrypt mes- sages while the designated receiver decrypts the ciphertext using the current private key. Hence, it is necessary for the users to update new private keys periodically. To revoke a user, the PKG simply stops providing the new private key for the user. It is obvious that a secure channel must be established between the PKG and each user to transmit the new private key and this would result in heavy load for the PKG.

In order to alleviate the load of the PKG in Boneh and Franklin's scheme, Boneh *et al.* [9] proposed another revocation method, called immediate revocation. Immediate revocation method employs a designated semi-trusted and online authority (i.e. mediator) to mitigate the management load of the PKG and assist users to decrypt ciphertext [10], [11], [12], [13]. In such a case, the online mediator must hold shares of all the users' private keys. Since the decryption operation must involve both parties, neither the user nor the online mediator can cheat one another. When a user was

revoked, the online mediator is instructed to stop assisting the user. However, the online mediator must help users to decrypt each ciphertext so that it becomes

a bottleneck for such schemes as the number of users On the other hand, in Boneh and Franklin's revocation method [2], all the users must periodically update new pri- vate keys sent by the PKG. As the number of users increases, the load of key updates becomes a bottleneck for the PKG. In 2008, Boldyreva *et al*. [14] proposed a revocable IBE scheme to improve the key update efficiency. Their revocable IBE scheme is based on the concept of the Fuzzy IBE [35] and adopts the complete subtree method to decrease the number of key updates from linear to logarithmic in the number private key size is 3log *n* points in an elliptic curve, where *n* is the number of leaf nodes (users) in the binary tree. (2) The scheme also results in enormous computation workload for encryption and decryption procedures. (3) It is enormous load for PKG to maintain the binary tree with a large amount of users.

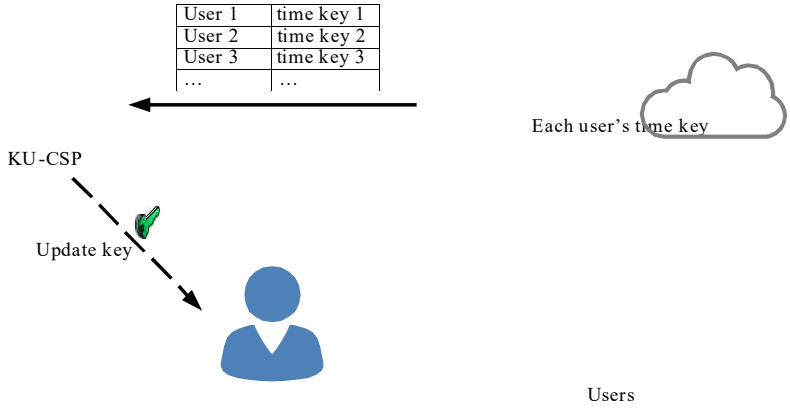


Fig. 1: Li *et al*.'s system model

Our Contributions

In order to solve both the un-scalability and the inefficiency in Li *et al*.'s scheme [24], we will propose a new revocable IBE scheme with cloud revocation authority (CRA). The proposed scheme possesses the advantages of both Tseng

and Tsai's revocable IBE scheme [23] and Li *et al*.'s scheme [24]. In particular, each user's private key still consists of an identity key and a time update key. We

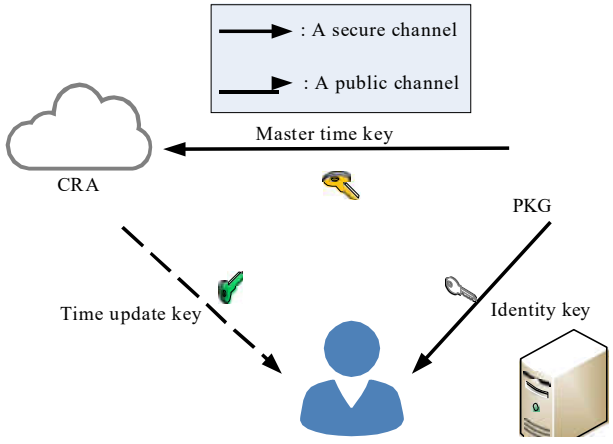


Fig. 2: System model for revocable IBE scheme with CRA

In this article, we first present the framework of our revocable IBE scheme with CRA and define its security notions to model possible threats and attacks. Accordingly, a new revocable IBE scheme with CRA is proposed. As the adversary model presented in [23], [24], it consists of two adversaries, namely, an inside adversary (or a revoked user) and an outside adversary. For security analysis, we formally demonstrate that our

grows enormously.

Moreover, Seo and Emura [17] refined the security model of Boldyreva *et al*.'s revocable IBE scheme [14] by consider- ing a new threat, called decryption key exposure attacks. Based on the idea of Libert and Vergnaud's scheme [16], they also proposed a revocable IBE scheme with decryption key exposure resistance. In order to reduce the sizes of both private keys and update keys, Park *et al*. [18] proposed a new revocable IBE scheme by using multilinear maps, but the size of the public parameters is dependent to the number of users. For achieving constant the size of the public parameters, Wang *et al*. [19] employed both the dual system encryption methodology [20] and the complete subtree method [14] to propose a new revocable IBE scheme.

introduce a cloud revocation authority (CRA) to replace the role of the KU-CSP in Li *et al*.'s scheme. The CRA only needs to hold a random secret value (master time key) for all the users without affecting the security of revocable IBE scheme. The CRA uses the master time key to generate the current time update key periodically for each non-revoked user and sends it to the user via a public channel. It is evident that our scheme solves the un-scalability problem of the KU-CSP. Our system model is depicted in Fig. 2.

scheme is semantically secure against adaptive-ID and chosen-ciphertext attacks (CCA) in the random oracle model under the bilinear decision Diffie-Hellman problem [2]. Finally, based on the proposed revocable IBE scheme with CRA, we construct a CRA-aided authentication scheme with period-limited privileges for managing a large number of various cloud services.

To demonstrate the merits of our scheme, Table 1 lists the comparisons among subtree-based IBE schemes [14], [16], [17], [18], [19], HIBE schemes [21], [22], Tseng-Tsai scheme [23], Li *et al.*'s scheme [24] and ours in terms of the usage of key update channel, the size of each user's private key, key update load, outsourced computation of authority, the workload of the PKG and scalability of authority.

Those subtree-based IBE schemes [14], [16], [17], [18], [19] and HIBE schemes [21], [22] employed the complete subtree method to decrease the number of key updates from linear to logarithmic in the number of users. However, each user's private key size is $O(\log n)$, where n is the number of users. These schemes still used a secure channel to transmit periodic private keys while no other authority shares the responsibility of user revocation. In Tseng and Tsai's revocable IBE scheme [23], both the identity key and time update key are issued by the PKG. In order to alleviate the load of the PKG, Li *et al.* [24] employed a key update cloud service provider (KU-CSP) to share the responsibility of user revocation. In our revocable IBE scheme, we em- ploy a cloud revocation authority (CRA) to perform user revocation. Indeed, the PKG in Li *et al.*'s scheme and ours may also perform the revocation operations. Both the KU- CSP and the CRA are designated to share responsibility for performing user revocation. For scalability, the KU-CSP in Li *et al.*'s scheme must keep n various time keys for n users so that it does not possess scalability and incurs the management load. On the contrast, the CRA in our scheme holds only one master time key for all the users. When the number n of users in the system is very large, the PKG may designate multiple CRAs to share the responsibility of user revocation while each CRA holds only the same master time key. However, in Li et al.'s scheme, each KU- CSP must also keep n time keys. Indeed, cloud computing is a ubiquitous computing environment so that putting mul- tiple CRAs on clouds may provide convenient management of user revocation while reducing the load of the single PKG. The detailed comparisons regarding computation and communication efficiency will be given in Section

(2) If ID does not appear in L_0, B chooses a random value $u \in Z_q^*$ and computes $S_{ID} = u \cdot P$. B adds $< ID, S_{ID}, u >$ in L_0 and returns $H_0(ID) = S_{ID}$ to A_I.

- *H_1-queries*: The challenger B maintains a list L_1 of tuples $< ID, i, T_{ID,i}, v, coin >$. Upon receiving the query along with (ID, i), B per- forms the following steps:

r secret key α and a master time key β. Finally, it sends β to the CRA via a secure channel. PP are made public to all the following algorithms.

- *Identity key extract* is a deterministic algorithm which is run by the PKG that takes as input the master secret key α and a user's identity ID, and outputs the corresponding identity key D_{ID}. Then, the PKG returns D_{ID} to the user via a secure channel.
- *Time key update* is a deterministic algorithm which is run by the CRA. The CRA uses the master time key β, a user's identity ID and a period i to compute the user's time update key $P_{ID,i}$ for period i. Then, the CRA returns the time update key $P_{ID,i}$ to the user via a public channel (e.g. e-mail or public board).

obtain a master secret key α, a master time key β and public parameters PP . It forwards PP to the adversary A_I while α and β are kept secret by B.

- *Phase* 1. The adversary A_I is allowed to issue the following queries in an adaptive manner.
  - *Identity key extract query* (ID). When A_I issues such a request along with a user's identity $ID \in \{0, 1\}^*$, B runs the *Identity key extract* algorithm to generate the identity key D_{ID} and sends it to A_I.
  - *Time key update query* (ID, i). When A_I issues such a request along with a user's identity $ID \in \{0, 1\}^*$ and a period i, B runs the *Time key update* algorithm to generate the time
- *ncryption* is probabilistic algorithm that is run by a user (sender). The sender takes as input a

- *System setup*. The challenger B first chooses a random master secret key $\alpha \in Z_q^*$ and sets $P_{pub} = \alpha P$. Then B provides A_I with $PP = < q, tt, tt_T, \hat{e}, P, P_{pub}, C_{pub}, H_0, H_1, H_2, H_3 >$, where $C_{pub} = aP$. Moreover, H_0, H_1, H_2 and H_3 are random oracles controlled by B defined as below.
  - *H_0-queries*: The challenger B maintains a list L_0 of tuples $< ID, S_{ID}, u >$. Upon receiving the query along with ID, B performs the following steps:
    (1) If ID appears in L_0, then B responds with $H_0(ID) = S_{ID}$.

    (1) If (ID, i) appears in L_1, then B responds with $H_1(ID, i) = T_{ID,i}$.

  H_3-queries: The challenger B maintains a list L_3 of pairs $< U, V, M, ID, i, w >$. Upon re- ceiving the query along with (U, V, M, ID, i), B performs the following steps:

  (1) If (U, V, M, ID, i) appears in L_3, then B responds with $H_3(U, V, M, ID, i) = w$.

(2) If (U, V, M, ID, i) does not appear in L_3, then B randomly chooses a string $w \in \{0,1\}^l$. B adds $< U, V, M, ID, i, w >$ in L_3 and returns $H_3(X) = w$ to A_I.

- *Phase* 1. A_I is able to issue three queries and B responds as follows.
  - *Identity key extract query* (ID): To respond to such a query, the challenger B first accesses the list L_0 to obtain u. Then, B sets the identity key $D_{ID} = u\,P_{pub}$ which is valid since the identity key $D_{ID} = u\,P_{pub} = u\,\alpha\,P = \alpha\,u\,P = \alpha\,S_{ID}$. B returns the identity key D_{ID} to A_I.
  - *Time key update query* (ID, i): To respond to such a query, B first accesses the list L_1 to obtain v and *coin*. If *coin* =1, B reports failure and terminates. If *coin* =0, B sets the time update key $P_{ID,i}=v\,C_{pub}$ which is valid since the time update key $P_{ID,i} = v\,C_{pub} = v\,aP = a\,v\,P = a\,T_{ID,i}$. B returns $P_{ID,i}$ to A_I.
  - *Decryption query* $(C = (U, V, W), ID, i)$: To respond to such a query, B first uses $(U, V, -, ID, i, W)$ to scan the list L_3 to obtain M. If $(U, V, -, ID, i, W)$ was not found, B returns failure and terminates which means that A can guess a right output value of H_3 hash function without using random oracles. Otherwise, B return M.
    (2) If (ID, i) does not appear in L_1, then B chooses a random value $v \in Z_q^*$. B then flips a random $coin \in \{0, 1\}$ and sets $T_{ID,i} = v\,P$ if $coin = 0$ and $T_{ID,i} = v\,bP$, otherwise. Finally, B adds $< ID, i, T_{ID,i}, v, coin >$ in L_1 and returns $H_1(ID, i) = T_{ID,i}$ to A_I. Indeed, when $coin = 1$, the third value bP of the DBDH problem is put in the corresponding query $H_1(ID, i)$. On the contrast, if $coin = 0$, the correspond- ing query $H_1(ID, i)$ does not include the DBDH problem. Note that the probability $\Pr[coin = 0]$ will be determined later. If we place the DBDH problem on every H_1 response, then the adversary cannot issue any *Timekeyupdate* query because the challenger cannot answer the correct time update key. In such a case, it cannot simulate the real adversary's ability.
  - H_2-*queries*: The challenger B maintains a list L_2 of pairs $< X, Y >$. Upon receiving the query along with X, B performs the following steps:
    (1) If X appears in L_2, then B responds with $H_2(X) = Y$.
    (2) If X does not appear in L_2, then B randomly chooses a string $Y \in \{0,1\}^{l}$. B adds $< X, Y >$ in L_2 and returns $H_2(X) = Y$ to A_I.
  - *Challenge*. In this phase, A_I issues two messages M_0, M_1, an identity ID^* and a period i^*. If ID^* does not appear in the list L_0, B randomly chooses $u^* \in Z_q^*$ and sets $S_{ID^*} = u^*\,P$. B adds the tuple $< ID^*, S_{ID^*}, u^* >$ in L_0. Meanwhile, B uses (ID^*, i^*) to scan the tuple $< ID^*, i^*, T_{ID^*,i^*}, v, coin >$ in the list L_1. If $coin = 0$, then B reports failure and terminates because (ID^*, i^*) is not the target identity and period. If $coin = 1$, B flips a random $\gamma \in \{0, 1\}$, receives Y^* by issuing the $H_2(\hat{e}(S_{ID^*}, \alpha\,cP) \oplus (v\,T))$ query and computes $V = M_\gamma \oplus Y^*$, where cP and T are the last two values of the DBDH problem. And B then selects a random string $W \in \{0,1\}^l$, adds the tuple $< U = cP, V = M_\gamma \oplus Y^*, M_\gamma, ID^*, i^*, W >$ in L_3. Finally, B returns the target ciphertext $C^* = (U, V, W = w)$ to A_I.
  - *Phase* 2. A_I may issue further queries as those in the *Phase* 1. The only restriction is that A_I cannot issue the *Time key update* query with (ID^*, i^*).
  - *Guess*. A_I outputs a guess γ'. The advantage ϵ of an IND-ID-CCA adversary A_I to attack the revocable IBE scheme with CRA is evaluated by the function $Adv_A = |\Pr[\gamma = \gamma'] - \frac{1}{2}|$. If the advantage ϵ of the adversary A_I is non-negligible, it means that the challenger B with non-negligible advantage can decide if $T = \hat{e}(P, P)^{abc}$. This resolves the DBDH problem with a non-negligible probability ϵ', which will be determined later.

Next, we analyze the probability that the simulation above will not abort. In the *Phases* 1 and 2, if *coin* = 0, the simulation continues. For convenience, let δ denote the probability that *coin* = 0. Since A_I makes at most q_u *Time key update* queries in the *Phases* 1 and 2, the probability that the simulation does not abort is δ^{q_u}. In the *Challenge* phase, if *coin*=1, the simulation continues, so the probability that the simulation does not abort is $1-\delta$. As a result, the total probability of the simulation not aborting is $\delta^{q_u}(1-\delta)$ in the *Phase* 1, *Phase* 2 and *Challenge* phases. By a similar technique in [28], we have that the maximum value of $\delta^{q_u}(1-\delta)$ is achieved at $\delta = 1 - 1/(q_u + 1)$ and so the probability of the simulation not aborting is at least $1/e(1 + q_u)$, where e is the base value of the natural logarithm. For handling the decryption query, if $(U, V, , ID, i, W)$ cannot be found in the list L_3, B returns failure and terminates, which means that A_I can guess a right output value of H_3 hash function. In this case, there are q_d *decryption queries*, the probability of A_I is at most q_d/q. In summary, B can solve the DBDH problem with probability $\epsilon' \geq \frac{\epsilon}{e(1+q_u)} - \frac{q_d}{q}$. Q

(2) If ID does not appear in L_0, B chooses a

ISBN 978-81-939386-1-4

random value $u \in Z_q^*$. B then flips a random $coin \in \{0, 1\}$ and sets $S_{ID} = u \cdot P$ if $coin = 0$ and $S_{ID} = u \cdot bP$, otherwise. Finally, B adds $< ID, S_{ID}, u, coin >$ in L_0 and returns $H_0(ID) = S_{ID}$ to A_{II}. Indeed, when $coin = 1$, the third value bP of the DBDH problem is put in the corresponding query $H_0(ID)$. On the con- trast, if $coin = 0$, the corresponding query $H_0(ID)$ does not include the DBDH prob- lem. Note that the probability $\Pr[coin = 0]$ will be determined later.

- H_1-*queries*: The challenger B maintains a list L_1 of public key $P_{pub} = \alpha \cdot P$. In addition, suppose that there are k independent service servers managed by k independent CRAs in the system. The PKG randomly selects k different master privilege keys $\beta_1, \beta_2, \ldots, \beta_k$ and sends each β_j to the correspond- ing CRA_j via a secure channel, respectively. In the meantime, the PKG also computes the privilege pub- lic key $C_{pub,j} = \beta_j \cdot P$ for each CRA_j. The PKG selects four hash functions $H_0, H_1 : \{0, 1\}^* \to tt$, $H_2 : tt_T \to \{0, 1\}^l$, $H_3 : \{0, 1\}^* \to \{0, 1\}^l$, where l is fixed. Finally, the PKG publishes the public parame- ters $PP = < q, tt, tt_T, \hat{e}, P, P_{pub}, C_{pub,1}, C_{pub,2}, \ldots, C_{pub,k}, H_0, H_1, H_2, H_3 >$.
- *Identity key extract*: As in the revocable IBE scheme with CRA proposed in Section 3. Upon receiving the identity $ID \in \{0, 1\}^*$ of a user, the PKG sends the identity key D_{ID} to the user via a secure channel.
- *Privilege key extract*: Suppose that a user with identity $ID \in \{0, 1\}^*$ is granted to have access to the service server j at period i. The corresponding CRA_j uses the master privilege key β_j to generate the period- limited privilege key $P_{ID,j,i} = \beta_j \cdot H_1(ID, i)$ and send it to the user via a public channel.
- *Authentication*: If a user with identity ID would like to access some service server j, the user sends an authentication request along with ID and period i to the service server j.
  - Upon receiving the authentication request, the service server j selects a tuples $< ID, i, T_{ID,i}, v > \in \{\}$. Upon receiv- ing the query along with (ID, i), B performs the following steps:
    1. If (ID, i) appears in L_1, then B responds with $H_1(ID, i) = T_{ID,i}$.
    2. If (ID, i) does not appear in L_1, then B chooses a random value $v \in Z_q^*$ and com- putes $T_{ID,i} = v \cdot P$. B adds $< ID, i, T_{ID,i}, v >$ in L_1 and returns $H_1(ID, i) = T_{ID,i}$ to A_{II}.

We say that the adversary F wins the IND-ID-CCA games (Games 1 and 2). This contradicts Theorem 3. Q

7 CONCLUSIONS

In this article, we proposed a new revocable IBE scheme with other authorities has been employed in Li *et al.*'s

requires higher computational and communicational costs than previously proposed IBE schemes. For the time key

challenge message $M \in \{0, 1\}^l$ and a random value $r \in Z_q^*$, and computes $U = r \cdot P$ and $V = M \cdot H_2((g_1 \cdot g_2)^r)$, where $g_1 = \hat{e}(S_{ID}, P_{pub}) = \hat{e}(H_0(ID), \alpha \cdot P)$ and $g_2 = \hat{e}(T_{ID,i}, C_{pub,j}) = \hat{e}(H_1(ID, i), \beta_j \cdot P)$. Finally, the service server j sends $C = (U, V)$ to the user.

- Upon receiving $C = (U, V)$, the user with the identity key D_{ID} and period-limited privilege key $P_{ID,j,i}$ computes $M = V \cdot H_2(\hat{e}(D_{ID} + P_{ID,j,i}, U))$. The user then sends $R = H_3(M, U, V, ID, i)$ to the service server j.
- Upon receiving the response message R, the service server j validates whether R is equal to $H_3(M, U, V, ID, i)$ or not. If so, the service server j accepts the request, and reject, other- wise.

Indeed, authentication (identification) schemes [42], [43], [44], [45] may be implemented by signature or encryption schemes. In our authentication procedure, the service server verifies a user by asking the user to decrypt a challenge ciphertext C. Then the user responds with R, which can pass the server's verification only when the user retrieves the valid plaintext M. The proposed CRA-aided authentication scheme with period-limited privileges aims at user identifi- cation and authorization before accessing service servers. The CRA-aided authentication scheme does not concern with the construction of secure session keys for encryption. Hence, some existing session key exchange protocol [46] or SSL protocol [47] can be employed to establish a secure session key for providing communication confidentiality

scheme with a cloud revocation authority (CRA), in which the revo- cation procedure is performed by the CRA to alleviate the load of the PKG. This outsourcing computation technique

revocable IBE scheme with KU-CSP. However, their scheme

In the following, based on the IND-ID-CCA security of the revocable IBE scheme with CRA, we prove that the proposed CRA-aided authentication scheme with period- limited privileges is secure under active attacks.

update procedure, the KU-CSP in Li *et al.*'s scheme must keep a secret value for each user so that it is lack of scalability. In our revocable IBE scheme with CRA, the CRA holds only a master time key to perform the time key update procedures for all the users without affecting security. As compared with Li *et al.*'s scheme, the performances of com- putation and communication are significantly improved. By experimental results and performance analysis, our scheme is well suited for mobile devices. For security analysis, we have demonstrated that our scheme is semantically secure against adaptive-ID attacks under the decisional bilinear Diffie-Hellman assumption. Finally, based on the proposed revocable IBE scheme with CRA, we constructed a CRA- aided authentication scheme with period-limited privileges for managing a large number of various cloud services.

ACKNOWLEDGMENTS

The authors would like to appreciate anonymous referees for their valuable comments and constructive suggestions. This research was partially supported by Ministry of Science and Technology, Taiwan, under contract no. MOST103-2221- E-022-MY2.

REFERENCES

[1] A. Shamir, "Identity-based cryptosystems and signature schemes," Proc. Crypto'84, LNCS, vol. 196, pp. 47-53, 1984.

[3] R. Housley, W. Polk, W. Ford, and D. Solo, "Internet X.509 public key infrastructure certificate and certificate revocation list (CRL) profile," IETF, RFC 3280, 2002.

[4] W. Aiello, S. Lodha, and R. Ostrovsky, "Fast digital identity revoca- tion," Proc. Crypto'98, LNCS, vol. 1462, pp. 137-152, 1998.

[5] M. Naor and K. Nissim, "Certificate revocation and certificate update," IEEE Journal on Selected Areas in Communications, vol. 18 , no. 4, pp. 561 - 570, 2000.

[6] S. Micali, "Novomodo: Scalable certificate validation and simplified PKI management," Proc. 1st Annual PKI Research Workshop, pp. 15-25, 2002.

[7] F. F. Elwailly, C. Gentry, and Z. Ramzan, "QuasiModo: Efficient certificate validation and revocation," Proc. PKC'04, LNCS, vol. 2947, pp. 375-388, 2004.

[8] V. Goyal, "Certificate revocation using fine grained certificate space partitioning," Proc. Financial Cryptography, LNCS, vol. 4886, pp. 247-259, 2007.

[9] D. Boneh, X. Ding, G. Tsudik, and C.-M. Wong, "A Method for fast revocation of public key certificates and security capabilities," Proc. 10th USENIX Security Symp., pp. 297-310. 2001.

[10] X. Ding and G. Tsudik, "Simple identity-based cryptography with mediated RSA," Proc. CT-RSA'03, LNCS, vol. 2612, pp. 193-210, 2003.

[11] B. Libert and J. J. Quisquater, "Efficient revocation and threshold pairing based cryptosystems," Proc. PODC2003, pp. 163-171, 2003.

[12] J. Baek and Y. Zheng, "Identity-based threshold decryption," Proc. PKC'04, LNCS, vol. 2947, pp. 262-276, 2004.

[13] H.-S. Ju, D.-Y. Kim, D.-H. Lee, H. Park, and K. Chun, "Modified ID-based threshold decryption and its application to mediated ID- based encryption," Proc. APWeb2006, LNCS, vol. 3841, pp. 720-725, 2006.

[14] A. Boldyreva, V. Goyal, and V. Kumar, "Identity-based encryption with efficient revocation," Proc. CCS'08, pp. 417-426, 2008.

[15] A. Sahai and B. Waters, "Fuzzy identity-based encryption," Proc. Eurocrypt'05, LNCS, vol. 3494, pp. 557-557, 2005.

[16] B. Libert and D. Vergnaud, "Adaptive-ID secure revocable identity-based encryption," Proc. CT-RSA'09, LNCS, vol. 5473, pp. 1-15, 2009.

[17] J.-H. Seo and K. Emura, "Revocable identity-based encryption revisited: security model and construction," Proc. PKC'13, LNCS, vol. 7778, pp. 216-234, 2013.

[18] S. Park, K. Lee, and D.H. Lee, "New constructions of revocable identity-based encryption from multilinear maps," IEEE Transac- tions on Information Forensics and Security, vol.10 , no. 8, pp. 1564
- 1577, 2015.

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

THE VEHICLE NETWORK ARCHITECTURE IN SMART CITY WITH BLOCKCHAIN TECHNOLOGY

Rajasekar Rangasamy* R.JAHNAVI **

*Professor, Department of Computer Science and Engineering,

** PG Scholar, Department of Computer Science and Engineering,

St. Peter's Engineering College, Hyderabad, Telangana, India.

ABSTRACT : *The* interconnected smart vehicle variety of sophisticated services to vehicle owners, car manufacturers and other providers of services. The rapid growth of the Internet of Vehicles (IoV) has brought the entire system enormous challenges for large data storage, intelligent management and information security. The traditional centralized management approach for IoV is facing the challenge of responding in real time. The block chain has already shown great advantages in the application of Bit coin as an efficient technology for decentralized distributed storage and security management. Block chain technology enables smart data –vehicles to be shared in a decentralized and manipulative manner while maintaining privacy, integrity, resilience and non-repudience. Smart vehicles, both consumers and producers of data smart cities, are increasingly interested. Vehicles can use intelligent city data to make decisions, such as dynamic traffic routing. In this paper, we examine how block chain innovation can be stretched out to vehicle organizing applications, considering the protected and dispersed storage of Big Data.

Keywords: Block chain Technology, Local Block Manger(LBM), IOV, OEMs(Original Equipment Manufacturer),Security, Vehicular Network Architecture, VANET, Smart Cities

INTRODUCTION

Smart cities are areas that support innovation through digital networks and applications (Figure 1). smart towns are often referred to as sustainable, digital or connected towns[1]. The aim of turning a city into an smart environment is to alleviate the problems of urbanization and an increasing urban population. A clever town is an urban area that offers sustainable economic growth and quality of life. Smart solutions, such as avoiding traffic congestion[2], Some of the technologies used are green buildings[3] and modern industrial control systems (ICS)[4] Could make urbanization sustainable today.

A clever city involves smart technology to Enhance how people live, work, travel. A key aspect of a clever city Vehicles of the next generation that incorporate new sensing, communication and social skills The broader concept of the Internet of Things. By providing wireless communication and mobile sensing, Vehicles can facilitate access to data, which is critical to smart cities a reality.

ISBN 978-81-939386-1-4

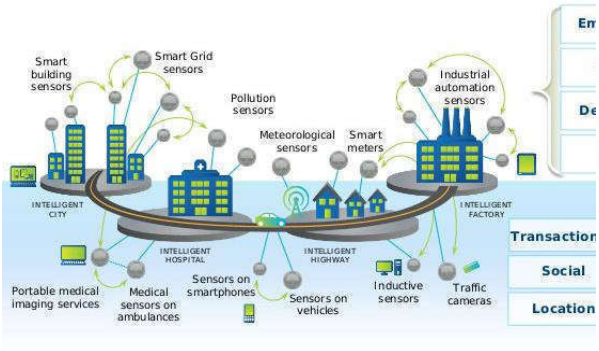


Fig.no: 1 SMART CITY

The characteristic advancement of the Internet (of PCs) to embedded and cyber physical frameworks that PCs themselves, albeit clearly not have computers Within a system of modest sensors and interconnected items, data on our reality and condition can be gathered at an a lot more elevated amount.Smart

vehicles are progressively associated with roadside framework, e.g. traffic the executives frameworks, to other close-by vehicles and, all the more for the most part, to the Internet, making vehicles part of the IoT.

Block chain[5] is a promising new technology(Figure 2) that can deal with the challenges mentioned above. It promotes decentralization, security and privacy and is widely used in various disciplines including finance, law, the IoT and energy[6].

Block chain has a Distributed block ledger that is shared across all nodes. Block chain maintains a distributed block library that is shared across all nodes involved. Every exchange, i.e. the trading of information between hubs, is confirmed by all or a portion of the hubs included, therefore dispensing with the requirement for focal experts.

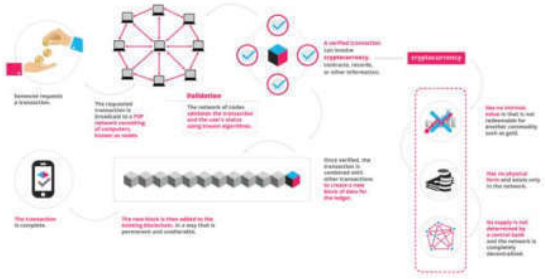


FIG.NO: 2 BLOCK CHAIN TECHNOLOGY

Exchanges are gathered together into squares and connected to the block chain. Adding another block to the block chain, known as mining, includes understanding a riddle or notwithstanding casting a ballot so as to achieve an agreement between the Peer-to-Peer (P2P) organize hubs that deal with the square chain distributively. Sensitive data transmitted between nodes that protect against eavesdropping can be encrypted. To ensure the security of the client, the Bitcoin cryptocurrency[7].

2. ADVANTAGES OF BLOCK CHAIN

2.1 Greater Transparency

The history of transactions becomes more transparent by using blockchain technology. Since blockchain is a dispersed record type, all members in the system share in distinguishable documentation from each duplicate. This common version can only be updated by consensus, which means that everyone has to agree.

To change a solitary exchange record, every consequent record would should be modified and the whole system would be associated. The blockchain information is in this manner increasingly precise, steady and straight forward Instead of pushing it through paper-substantial procedures.It is also open to all participants who have access.2.2 Enhanced Security **There are a number of ways in which blockchain is safer than other recording systems. Transactions must be agreed in advance of recording. Following endorsement of the exchange, it is encoded and connected to the past exchange. This makes it extremely trouble some for programmers to bargain exchange information, together with the way that data is put away on a system of PCs rather than on a solitary server. In any industry where the insurance of delicate information is critical money related administrations, government, social insurance—blockchain truly has the chance to change the way critical information is shared by preventing fraud and unauthorized activity.**

ISBN 978-81-939386-1-4

2.3 Improved Traceability

On the off chance that your organization manages items that are exchanged a mind boggling store network, you realize that it is so hard to follow an item back to its starting point. At the point when ware trades are recorded on a blockchain, you wind up with a review trail demonstrating where a benefit originated from and where it stops made on its adventure. This historical transaction data can help check asset authenticity and prevent fraud.

2.4 Increased Efficiency And Speed (I should start here)

When using traditional, paper-intensive processes, trade is a time consuming process that is prone to human error and often requires mediation from third parties. Exchanges can be finished speedier and all the more effectively by streamlining and robotizing these procedures with a block chain. Since recording is carried out using a single digital booklet shared between participants, you do not have to reconcile multiple booklets and end up with less clutter. Furthermore, it winds up less demanding to confide in one another without the requirement for various middle people when everybody approaches a similar data. Clearing and settlement can therefore occur much faster.

2.5 Reduced Costs

Decreasing expenses is a need for generally organizations. You don't require the same number of outsiders or delegates to make ensures with blockchain on the grounds that it doesn't make a difference whether you can believe your exchanging accomplice. Instead, you just have to confide the blockchain data. You will also not need to review so much documentation to complete a trade, because everyone has access to a single, unchanging version.

3. ARCHITECTURE OF SMART VEHICLES

Our BC-based architecture described in this area can satisfy the up to referenced security, protection and versatility prerequisites. The overlay, as appeared in (Figure 3), is involved unique hubs including keen vehicles, smart structures (that can be keen homes or administration focuses), OEMs(Original Equipment Manufacturer), vehicle mechanical production systems, SW suppliers, cloud stockpiles, and client's cell phones such as cell phones, workstations, or tablets. Each shrewd building is halfway overseen by a believed hub known as Local Block Manager (LBM).

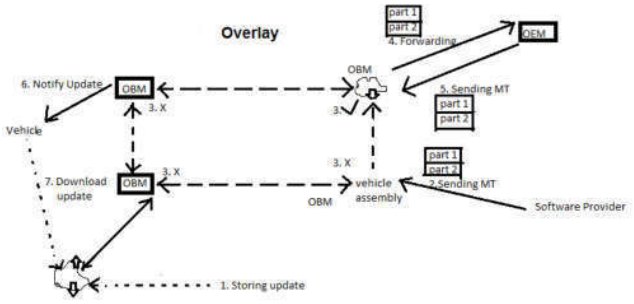


Fig.no: 3 ARCHITECTURE OF SMART VEHICLES

In the proposed design, a vehicle can interface with the overlay either straightforwardly or through the smart building. A vehicle is a piece of a smart building while it is inside the scope of the LBM (e.g., while it is stopped at home). Each smart building has a safe local storage. In an administration focus this storage will be used to gather vehicles-explicit information (e.g., the date of the last vehicle upkeep) and additionally data required by the mechanics to keep up the vehicle (e.g., the most recent accessible SW updates and fix directions). In a smart home, vehicle related information, client information (e.g., originating from brilliant gadgets, for example, cell phones), and information originating from neighborhood sensors and actuators can be gathered in the local storage

A smart vehicle can likewise specifically associate with the overlay arrange.
There are two types of cloud storages in the overlay. There are
The first category is overseen by the OEM or potentially SW refresh supplier and could be

utilized to store most recent SW refreshes. The second gathering incorporates storage which vehicle proprietors use for putting away vehicular information to get explicit administrations, e.g. cloud capacity given by vehicle insurance agencies.
The overlay is a clustered network and its groups are kept up by Cluster Heads (CH). The CHs likewise deal with the open BC by checking and putting away exchanges and squares and are thus likewise alluded to as Overlay Block Manager (OBM).

The OBMs are picked by the cluster individuals utilizing techniques, for example, portrayed in [9]. The group individuals can choose another CH on the off chance that an OBM fizzles or is imperiled. In the overlay there are two kinds of exchanges, specifically:
1. Single signature transactions
2. Multi sig transactions

In the vehicle gathering the OEM will store its PK on each amassed vehicle and the vehicle will produce a mystery key combine (e.g., a RSA key match comprising of a private and an open key). Both the PK of the OEM and the key match will be se-curely put away on the WVI in a carefully designed capacity. While a vehicle is gathered it can likewise make a beginning exchange, an underlying exchange required to take part in the BC.

Overlay exchanges are communicated and checked by the OBMs. An OBM checks an exchange by confirming the mark of the exchange members with their PK. Furthermore, the OBM confirms if the past exchange referenced in the second field of every exchange exists in people in general BC and is legitimate. OBMs create new squares and communicate them to other OBMs.

On accepting another square, an OBM confirms the exchanges inside the square before adding it to its own BC and communicate it to different OBMs. A versatile vehicle that changes its group can still speak with other overlay hubs and store exchanges on the general population BC, yet it will be unavailable for overlay hubs as the vehicle PK isn't in the key lists of the new OBM. To make itself open, the vehicle should store its key on the key lists of its new OBM.

4. SMART CONTRACT & SUPPLY CHAINUSING BLOCK CHAIN

Blockchain based on smart contracts is expected to play an important role in the management, control and most importantly the securing of IoT devices in the context of IoT.

Nick Szabo first authored the term Smart Contracts in 1994. (Figure 4) A smart contract is on an exceptionally essential dimension an electronic exchange custom that executes the terms of the assention. In the oversimplified definition, smart contracts are programs composed by clients to be transferred and executed on the blockchain. Smart contracts empower you to exchange money, property, offers or anything of huge worth in a direct, conflict - free way while avoiding a specialist 's organizations. The most ideal approach to depict Smart contracts is by contrasting innovation and a merchant.

You would typically go to an attorney or a public accountant and pay them and hang tight for the report. You simply drop a bitcoin into the moving machine with smart contracts (i.e. Record), and your escrow, driver's permit, or any drops to your record.

All the more along these lines, smart contracts not just characterize the guidelines and punishments around an understanding similarly that a customary contract does, yet additionally consequently implement those commitments.

In block chain solidity, which is a JavaScript-like language, is the script or programming language for intelligent contracts. Ethereum Blockchain supplies EVM (Ethereum Virtual Machines) that are essentially the mineral nodes. These nodes are able to carry out and enforce these programs or contracts in a cryptographically reliable manner.

Ethereum supports Ether's own digital currency. As in Bitcoin in Ethereum, users can transfer coins to each other using normal transactions recorded on the booklet and there is no need for a blockchain status in Bitcoin for such transactions.

An intelligent contract has its own account and address and its own executable code and balance of Ether coins is associated with it.EVM storage is moderately costly and another remote, decentralized information store, for example, Bit Torrent, IPFS or Swarm can be utilized to transfer expansive capacity to the blockchain.

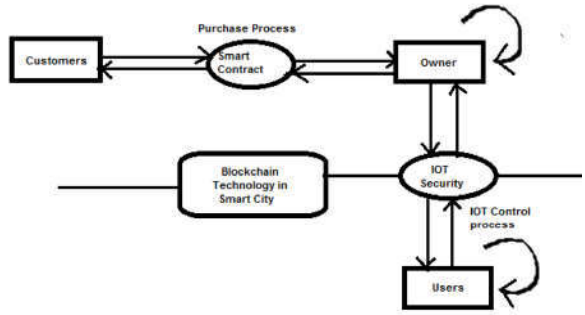


FIG.NO:4 SMART CONTRACTS

5. CHALLENGES IN VEHICULAR NETWORK

5.1Technical Challenges:

The highlights of VANETs additionally assume an essential job in the transmission of the packets. The transmission challenges identified during the packet transmission include next hop selection, queuing disciplines and duration of the paths. DSR / GPSR Protocols maintain neighborhood lists for next hop determination. In the event that the rundowns are not right, the best next jump could be missed or more terrible, a vehicle hub that is as of now out of the scope of transmission could be chosen. The maintenance of updated lists requires the frequent broadcast of " hello " packets between the network nodes.

5.2 Security Challenges:

Security has so far gotten less consideration among every one of the difficulties of the VANET. VANET packets may contain life - critical information and it is therefore necessary to ensure that these packets are not attacked by the intruder. Likewise, drivers ' liability is established by informing them of the traffic environment correctly at specified time.

5.3 Social and Economic Challenges:

In addition to the technical challenges involved in the use of the VANET, there are social and economic challenges. It is difficult to build a system that transmits a traffic signal violation that can increase production costs that the manufacturer cannot deal with.

6. SECURITY IN VEHICULAR NETWORK

6.1 Inspections:

In most of the United States States, inspection of all vehicles must pass once a year. This annual trip to the mechanic offers interesting security maintenance options as well as typical maintenance. The inspection process gives the system a chance to return to a known basic state and acts as a firebreak against worms and virus infections. These inspections can also be included in the standard maintenance package whenever a driver enters the vehicle for a tune – up.

6.2 Honest Majority:

Another benefit of vehicle networks is that most drivers will probably be honest. This is reinforced by the fact that few people feel comfortable with their vehicles, so that most drivers simply accept the default settings. If vehicles have adequate safeguards2 against worms and viruses, drivers ' trustworthiness becomes vehicles that correctly adhere to established protocols.

6.3 Registration central:

In contrast to ad hoc networks, all nodes in a vehicle network must be registered with a central authority, such as the state (in the US).These institutions have already developed extensive infrastructure to track and manage these registrations, and in the long term, we could use this existing bureaucracy to improve the safety of vehicle applications. However, given the vast nature of the existing infrastructure, it takes years of effort to make changes and requires considerable funding.

6.4 Access Controlled:

There are already access control mechanisms for many parts of the transport system. For instance, toll streets and numerous bridges

have controlled passage and leave focuses, so it might be down to earth to furnish these areas with the vital foundation to distribute ephemeral identities and/or keys to enable vehicles to communicate while on the controlled road.

7. APPLICATIONS

7.1 Authenticated Message Origin Location:

The location will play an important role in many vehicle applications, making it essential to determine whether a message originated at a particular location. It would also prevent an opponent from using another communication medium (e.g. cellular) to replay a message heard in one location as if it originated in another location, thus preventing the types of attacks in sensor networks by wormholes[8].

7.2 Traffic Congestion Detection

In this application, vehicles recognize when the quantity of neighboring vehicles surpasses a certain limit (and additionally their normal speed dips under a given edge), and after that hand-off this tally to vehicles moving toward the area of the congestion.

7.3 Warning System For Deceleration

In this application, if a vehicle significantly reduces its speed,1 A warning will be broadcast along to other vehicles with its current position and speed. Beneficiaries will communicate the message to vehicles and any vehicles their d will be alerted directly behind the vehicle in question Alert their drivers by means of visual and/or audio signals.

8. Conclusion

In this paper, we proposed a BC-based security architecture for smart connected vehicles capable of supporting a wide range of (future) automotive applications. We evaluated our architecture utilizing a proof-of-idea usage of a remote SW refresh framework and utilize the last to demonstrate the relevance of our design and also its advantages contrasted with a declaration based framework.
The privacy of users is guaranteed through the use of changing public keys (PK). The security of our architecture is largely due to the strong security features of the BC technology underlying it. The OBMs also provide access control for transactions sent to the members of their clusters.
The architecture supports emerging automotive services by providing a safe and reliable way to exchange data while safeguarding the safety of the end user.

9. REFERENCES

[1.] Xiong, Z.; Sheng, H.; Rong, W.; Cooper, D.E. Intelligent transportation systems for smart cities: A progress
review. Sci. China Inf. Sci. 2012, 55, 2908–2914.
[2.] Bowerman, B.; Braverman, J.; Taylor, J.; Todosow, H.; von Wimmersperg, U. The vision of a smart city.In Proceedings of the 2nd International Life Extension TechnologyWorkshop, Paris, France, 28 September
2000; Volume 28.
[3.] Su, K.; Li, J.; Fu, H. Smart city and the applications. In Proceedings of the 2011 International Conference on
Electronics, Communications and Control (ICECC), Zhejiang, China, 9–11 September 2011; pp. 1028–1031.
[4.] Hollands, R.G. Will the real smart city please stand up? Intelligent, progressive or entrepreneurial? City
2008, 12, 303–320.
[5.] A. Dorri, S. S. Kanhere, and R. Jurdak, "Towards an optimized blockchain for iot," in 2nd International Conference on Internet-of-Things Design and Implementation. ACM, 2017, pp. 173–178.
[6.] F. Tschorsch and B. Scheuermann, "Bitcoin and beyond: A technical survey on decentralized digital currencies," IEEE Communications Surveys Tutorials, vol. 18, no. 3, pp. 2084–2123, 2016.
[7.] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system, http://bitcoin.org/bitcoin.pdf, retrieved 25 jun 2018," 2008.
[8.] Yih-Chun Hu, Adrian Perrig, and David B. Johnson. Packet leashes: A defense against wormhole attacks in wireless networks. In Proc. of IEEE INFOCOM, April 2003.
[9.] Hancke, G.; Silva, B.; Hancke, G., Jr. The Role of Advanced Sensing in Smart Cities. Sensors 2012, 13, 393–425.

[CrossRef] [PubMed]
[10.] Kyriazis, D.; Varvarigou, T.; White, D.; Rossi, A.; Cooper, J. Sustainable smart city IoT applications: Heat and electricity management amp; Eco-conscious cruise control for public transportation. In Proceedings of the 2013 IEEE 14th International Symposium on "AWorld of Wireless, Mobile and Multimedia Networks" (WoWMoM), Madrid, Spain, 4–7 June 2013; pp. 1–5.
[11.] Zanella, A.; Bui, N.; Castellani, A.; Vangelista, L.; Zorzi, M. Internet of Things for Smart Cities. IEEE Internet
Things J. 2014, 1, 22–32.
[12.] Elmangoush, A.; Alhazmi, A.; Magedanz, T.; Schuch,W.; Estevez, C.; Ehijo, A.;Wu, J.; Nguyen, T.; Ventura, N.;Mwangama, J.; et al. Towards Unified Smart City Communication Platforms. In Proceedings of theWorkshop on Research in Information Systems and Technologies, Chillán, Chile, 16 October 2015.
[13.] Satoshi Nakamoto, "Bitcoin: A peer-to-peer electronic cash system ",2008, https://s3. Amazon aws. com/academia.edu.documents/32413652/BitCoin_P2P_electronic_cash_system.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1506052605&Signature=ua%2FsjZRbx0wc7Tx35Kl1dwUwmqo%3D&responsecontentdisposition=inline%3B%20filename%3DBitcoin_A_Peer-toPeer_Electronic_Cash_S.pdf, accessed on Sep 22, 2017.
[14.] Inspiring, http://www.inspiring.co.uk, accessed on Sep 20, 2017.
[15.] Slock, https://slock.it, accessed on Sep 20, 2017.
[16.] J.-P. Hubaux, S. Capkun, and J. Luo, "The Security and Privacy of Smart Vehicles," IEEE Security and Privacy Mag., vol. 2, no. 3, May–June 2004, pp. 49–55.
[17.] B. Parno and A. Perrig, "Challenges in Securing Vehicular Networks," Wksp. Hot Topics in Networks (HotNetsIV), 2005.
[18.] Jean-Pierre Hubaux, Srdan ¯ Capkun, and Jun Luo. The ˇ security and privacy of smart vehicles. IEEE Security & Privacy Magazine, 2(3):49–55, May–June 2004.
[19.]T. Bakıcı, E. Almirall, J. Wareham, A smart city initiative: the case of Barcelona, J. Knowl.Econ.(2012),http://dx.doi.org/10.1007/s13132-012-0084-9(Availableat
http://www.springerlink.com/content/9318pq8q61r06345/

Proceedings on 4th International Conference on Emerging Technologies in Computer Science(ICETCS-2018)
ISBN 978-81-939386-1-4

A Survey on Detection and Prediction of Lung Cancer Using Different Algorithms

[1]N. Sudhir Reddy, [2]Dr.V Khanna,
[1]Research Scholar, [2]Professor, Dean Information
[12]Department of Computer Science and Engineering
[12]Bharath Institute of Higher Education and Research (BIHER), Chennai, Tamil Nadu
E-mail: sudhir.navar@mrce.in

***Abstract*: One of the major causes of cancer death is through Lung cancer. The overlapping of cancer cells acts as an impediment for its early detection. Identifying genetic and environmental factors plays a vital role in developing better techniques for its prevention. And in order to discover the anomalies in target images, time factor is paramount. In this work, decision tree algorithm is used for prediction of lung cancer wherein the important pattern with their corresponding weightage and score is studied. Processes such as preprocessing of images and feature extraction are done using Histogram Equalization and using neural network classifier does normal or abnormality check of the patient. If lung cancer is detected early, then it will reduce many treatment options, reduce risk of invasive surgery and increase survival rate. Therefore, an easy, cost effective and time saving method will produce promising result for detection and prediction of lung cancer.**

Keywords: ***Data Mining, Decision Tree Algorithm, Early Detection, Histogram Equalization, Image Processing, Lung Cancer, Prediction.***

I. INTRODUCTION

To enhance the manual analysis, Image processing techniques acts as a good quality tool. In order to enhance the images during early detection and treatment stages, the Image processing technique is used in this proposed method. Assessment of Image quality and its improvement depends on the enhancement stage. In this stage, the pre-processing technique is used based on principal component analysis and Histogram Equalization. During the digital image analysis, classification is very important as it categorizes images in to groups as per their similarities. In proposed system, Histogram Equalization is used for preprocessing of images and feature extraction are done using Histogram Equalization and using neural network classifier does normal or abnormality check of the patient. Subsequently, survival rate of patient is predicted through the extracted features.

Many significant patterns such as Smoking, Environment, Alcohol, Obesity, Chronic Lung Disease, Balance Diet, and Mental trauma, Radiation Therapy, Tobacco and Genetic Risk are considered in this proposed early detection and prediction system. One of the most common causes (about 90%) of lung cancer is due to cigarette smoking as cigarette smoke contains more than 4,000 chemicals, many of which have been identified as carcinogenic. Other factors that contributes to lung cancer are environment pollution mainly air and excessive alcohol consumption. The risk of developing lung cancer is 20-25 times greater for someone whose per day consumption of cigarettes exceeds one pack than someone who abstains from smoking. Lung cancer begins in one or both lungs and occurs for out-of-control cell growth. There are other problems if Lung cancer extends to the brain affecting vision and weakness on one side of the body. Cough, coughing up blood, chest pain and shortness of breath are signs of primary lung cancers.

Chest Radiograph (x-ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI scan) and Sputum Cytology are some of the advanced techniques for diagnosing lung cancer. However, these techniques are not affordable for many and also time consuming. A new technique to diagnose the lung cancer in its early stages is the need of the hour as most of the aforementioned techniques are capable of detecting the lung cancer only in its advanced stages, reducing the patient's chance of survival. Hence, manual analysis can be enhanced through Image processing techniques.

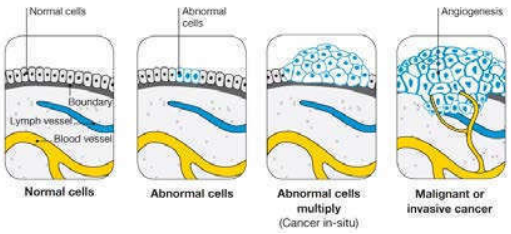


Fig.1: The beginning of cancer

This lung cancer risk prediction system using the significant pattern prediction tools will enhance the detection process. And thus, early prediction should play a significant role in the diagnosis process and for an effective preventive strategy.

As mentioned earlier, the existing techniques to diagnose lung cancer are not affordable for many and also time consuming. Also, it detects the cancer in its advanced stages ultimately reducing patients' chance of survival.

Therefore, the proposed system is designed to predict the lung cancer in its early stage depending

ISBN 978-81-939386-1-4

on few factors and thresholding. Here, the time and cost required for excessive medical test is reduced as the number of rule for testing in this system is lesser. Another advantage is that, the proposed system is web based connecting patients from remote locations to directly communicate with doctors.

Furnished below are the key objectives of this system:

- To bypass some testing rules not necessarily required.
- To make the system more time and cost efficient.
- To enhance the performance accuracy of the System.
- To use lesser number of attributes
- To detect during the initial stage.
- To increase the survival rate of the patient by more than 5 years.

The literature survey is conducted extensively in order to acquire enough relevant information concerning to topic of research.

II. LITERATURE REVIEW

The aim of most of the earlier work is to develop an efficient system to predict cancer in its early stages.

In this paper, it was intended to study the different characteristics of information mining procedures as data processing ideas benefits in carcinoma classification. For applicable rule generation and classifications on diseases, it studies different data processing and hymenopteron colony improvement techniques. Furthermore, it provides the fundamental framework for enhancement in cancer diagnosis..

In this work, the preprocessing of the photographs is done using bar chart equalization. In order to decide the performance of the various classifiers, experimental analysis is formed with dataset wherein performance depends on the correct and incorrect classification of the classifier. Neural Network rule is implemented and performance comparison with other classification algorithms is done. It yields the highest TP Rate and lowest FP Rate clearly proving its efficiency as compare to alternative classifiers. Subsequent work from the same author relies on Feature Extraction and Principal element Analysis for detection of carcinoma in CT scan pictures. It utilizes a hybrid technique supported feature extraction and Principal element Analysis (PCA) [1].

The image quality and accuracy forms the key factors of the analysis. With the help of segmentation principles, associate degree increased region of the thing of interest is used as a basis for feature extraction. A normality comparison is formed by counting on general options. [4].

The primary target of the work is vital improvement in distinction of plenty with the suppression of background tissues. It is obtained by standardizing the parameters of the planned transformation. The manual analysis of the phlegm samples is not accurate, time taking and requires trained person to avert any errors during diagnosis. The segmentation results used as a base for a pc motor-assisted designation system for early detection of cancer enhances the probabilities of survival for the patient. In this work, physicist filter is used which enhances the medical pictures. [5].

In order to divide phlegm color pictures and detect carcinoma during early stages, this paper presents Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) bunch rule. The segmentation results are used as a base for a pc motor-assisted designation (CAD) system for early detection of carcinoma which can improve the probabilities of survival for the patient. But segmentation results are inaccurate due to variation within the grey level and also the relative distinction among the photographs. Therefore, a thresholding technique is utilized as a pre-processing step. Subsequently, the quantitative procedures area unit supported the nuclear feature and proved the thresholding rule to be successful in extracting the nuclei and protoplasm regions. A thousand phlegm color pictures were used in this study to check each strategy. HNN proved to be more efficient in classification compared to FCM as it extracted the nuclei and protoplasm regions. An accuracy of 98% proves the efficiency of the thresholding classifier [2].

3. Steps involved

The following steps are involved in the diagnosis of carcinoma:

Images captured

Preprocessing of images

Image segmentation

Feature extraction

Principle element analysis

Neural network classifier

Diagnosis result

Prediction method

Images captured or knowledge assortment

CT-Scan pictures of cancer and non-cancer patients are gathered. Usually it is taken from different diagnostic centers that have no connection. The digitized pictures area unit keeps within the DIACOM format with a resolution of eight bits per plane [1].

Preprocessing of images

With an aim to improve the interpretability or sensitivity of knowledge enclosed, image Pre-processing stage during this system commences with image improvement. This is done in order to provide higher input for alternative programmed image process techniques. Abstraction domain strategies and frequency domain strategies are the two major categories into which Image improvement techniques are divided. Subsequently for alternative image process, the image improvement techniques area unit is utilized as pre-process tools and the quantitative measures will validate that techniques area unit are most fitted.

In the pre-processing stage, the initial image is not disturbed but there is selective elimination of the redundancy in scanned pictures. This is very important within the designation of carcinoma. To strengthen the superiority of every image, they are preprocessed with the Histogram- equalization.

III. IMAGE SEGMENTATION

S. No	Study of	No. of sample images	Method used	Performance Results
1.	Armato et al.[2]	24 datasets	Iterative thresholding and morphological operations	RmsD=0.54 (0.8 pixel)
2.	Yim et al. [26]	10 subjects	Region growing, connected component	RmsD = 1.2 pixel
3.	Pu et al. [19]	20 datasets	Thresholding	FP/GT = 0.43%, FN/GT = 1.63%
4.	Gao et al. [8]	8 subjects	Thresholding	DSC = 0.9946
5.	John et al. [12]	60 slices	Multilevel thresholding	-
6.	Itai et al [11]	-	Snake deformable model	Qualitative assessment
7.	Silveria et al. [21]	1 slice	Level set Active contour	Qualitative assessment
8.	Annangi et al. [1]	1130 image	Shape-based deformable Model	DSC = 0.88
9.	Shi et al. [20]	247 image	Shape-based model and PCA	OM = 0.92, AD = 1.78 pixel
10.	Sun et al. [23]	30 scans	Active Shape Models and optimal surface finding methods	DSC = 0.975
11.	Besbes and Paragios [3]	247 image	Shape-based deformable Model with boosted image cues	OM = 0.9474, AD = 1.39 pixel
12.	Sofka et al. [22]	260 scans	Graph-based shape with iterative surface deformation	SCD = 1.95
13.	Hua et al. [10]	15 scans	Graph-search	Sen. = 0.986, Spec. = 0.995
14.	Kockelkorn et al. [14]	22 scans	Prior training, statistical Classifier using LII	OM = 0.96, AD = 1.68mm
15.	El-Ba et al. [7]	10 image datasets	Statistical MGRF model	Accu. = 0.968
17.	Dai et al. [6]	23 scans	2D wavelet transform	RmsD = 0.52mm
18.	Yankelevitz et al. [25]	Not specified	Graph-cuts and GMM	DSC = 0.984

IV. FEATURE EXTRACTION

In order to find and segregate the most well liked parts or shapes of an inputted image, algorithms associate degreed strategies are used during image options Extraction stage. The binarization and GLCM are the two strategies to predict the presence of carcinoma. The respective strategies area unit supported facts that powerfully associated with respiratory organ anatomy and knowledge of respiratory organ CT imaging.

A. Binarization Approach

The binarization approach is employed for detection of cancer. In order to validate whether the cells are normal or abnormal in the respiratory organ, this approach extracts the amount of white pixels and check them against some threshold. Subsequently, validation is done on whether a range of white pixels of a brand new image is a smaller amount than the edge. Depending on the quantity of white pixels, it indicates whether the image is normal or if the quantity of the white pixels is larger than the edge, it specifies that the image in abnormal.

B. GLCM (Grey Level Co-occurrence Method)

The gray levels in a picture (that is the brightness value) can be tabulated using GLCM. The gray-level co-occurrence matrix from image in MATLAB is formed as the beginning of this process.

In the subsequent stage, the GLCM is normalized

Where: i is that the row range and

J: is that the column range

And thus the texture events from the GLCM can be found [1].

C. Masking Approach

This approach has high dependency on numerous area unit that looks like white connected areas within ROI (lungs). With the increase in the

percentage of the area unit inside lungs, the cancer increases. With the integration of the aforementioned two approaches, it is concluded that if the range of black pixels in the image is bigger than white pixels then that image is traditional. But if the range of black pixels in the image is smaller than white pixels then that image is considered to be having an anomaly.

D. PCA (Principle element Analysis)

The information in image can be regularized with the help of this technique. Generally, in Real-world knowledge sets the variables exhibit some associations. These kind of area unit associations are either linear or a minimum making them feasible for common examination techniques. Principal element analysis is one such method that rotates the initial knowledge to new coordinates. The information thereby becomes uniform. [1].

PCA comprises of the following processes:

- The mean and variance of the options within the image are calculated.
- The sample mean are subtracted from all the observations.
- Then divided by the sample variance that scales and centers the information.
- The coefficients of the key components and their applicable changes area unit are then calculated by finding the Manfred Eigen operate of the sample variance matrix.
- For the key components, the matrix holds the coefficients and diagonal parts store the modification of the relevant principal constituents.
- The diagonal is mined.
- The maximum info content is arrived through the differences in knowledge that is necessary for higher classification [1].

V. NEURAL NETWORK CLASSIFIER

The tool used here is supervised feed-forward back-propagation neural network ensemble. This tool differs in many aspects from other classifiers like theorem and k – nearest neighbor classifiers. In the later two classifiers, the linear knowledge is necessitated to accurately figure. But neural network operates for non- linear knowledge because of which it is simulated on the reflection of biological neurons and network of neurons.

The detection accuracy can be enhanced by training the neural network with broader selection of input file. This would mean that the system might be influenced with a tiny low set {of knowledge of knowledge of information} or massive set of comparable data. Neural network classifier requires an outsized set of information for coaching. But coaching together is time consuming that is an impediment to attain stability. So training them as biological neural network transmitting signals at a speed equivalent to electrical signals can enhance the speed drastically.

Neural network comprises of three layers namely the Input layer, internal hidden layer and output layer area unit. In the input layer, the nodes of area unit are coupled with variety of nodes within the internal hidden layer. Linking of nodes within the internal hidden layer might take place either with another internal hidden layer or with associate degree output layer. Further, the output layer comprises of one or more response variables.

The final Steps comprises of the following in Neural Network Classifier:-

- Feed-forward back propagation network is created.
- The accessible samples and cluster outlined for it are trained with neural network.

From the results of neural network, the area unit is determined. The samples are then trained in network classification rate and calculated using mathematical formulas.

A. Support Vector Machines (SVMs)

For the purpose of classification, the algorithms such as support vector machines area unit supervised learning models with associated learning algorithms that examine knowledge and its patterns are used. Assuming that the pattern classifier using a hyper plane to segregate two types of patterns supported given examples i=1n .Where (i) could be a vector within the input area I=Rk and y (i) denotes the category index taking worth one or zero.

Binary categories are categorized by support vector machine. They find and employ a class boundary (the hyper plane) increasing the margin within the given coaching knowledge. The support vectors are the coaching knowledge samples on the hyper planes close to the category boundary area unit. And the distance between the support vectors and the category boundary hyper planes is known as margins. The SVM area unit supported the idea of call planes that outline call boundaries.

The below representation shows the Theory of SVM:

Given coaching knowledge (xi ,yi) for i = one ...N, with xi, Rd and Lolo ->

Every input x is ab initio mapped into the next dimension feature area F, by $z=\varphi(x)$ through a nonlinear mapping φ:

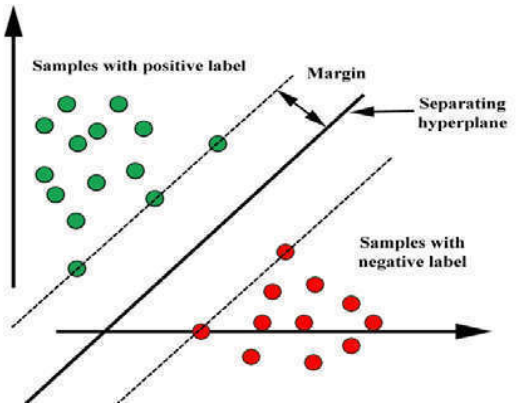
$Rn \rightarrow F$.

$f(x) = WTx + b$

W is that the traditional to the road, and b the bias. W is thought because the weight vector.

Assets of objects having completely different category memberships are segregated by decision

plane. For knowledge classification, SVM can be considered as a very valuable method. In coaching and testing knowledge comprising of some knowledge instances, classification task is generally involved. One target value/class labels and numerous other attributes/features are an integral part of every instance within the coaching set.



VI. Diagnosis Result

S. No	Study of	Parameters used	Method used	Performance Results (Area under ROC curve)
1	Kawata, Yoshiki et al. [13]	Curvature Index	Linear Discriminant classifier	*Az* =0.94
2	McNitt-Gray et al. [17]	shape, size, attenuation, and texture	Stepwise discriminant Analysis	Accuracy = 90.3%
3	Cristobal et al. [5]	nodule border smoothness, speculation, and	Neural network classifier	*Az* = 0..871
4	Gurney and Swensen [62]	nodule border smoothness, speculation, and	Bayesian classifier	*Az* = 0.894
5	Matsuki, Yuichi et al. [16]	Shape – based Features	Neural network classifier	*Az* = 0.951
6	Henschke et al. [9]	-	Statistical-multiple-object detection and location system (S-MODALS)	-
7	Lo et al. [15]	Radiographical parameters	Neural network classifier	*Az* = 0.89
8	Suzuki et al. [23]	-	MTANN classifier	*Az* = 0.89
9	Chen et al. [4]	-	ANN ensemble	*Az* = 0.91
10	Nakurama et al. [18]	Subjective Features	ural network classifier	*Az* = 0.854

VII. Prediction Method

Cancer at a very advanced stage does not have a remedy and death becomes inevitable. Therefore, detecting the carcinoma at an early stage plays a vital role within the designation method. In this paper, an efficient carcinoma prediction system that supports data processing is planned. The risk prediction system proposed in this work is expected to enhance the detection process. Many risk factors such as genetic, age, diet, gender, medical history, sedentary lifestyle, obesity, environment, mental trauma, heart condition, consumption of excessive alcohol, tobacco and anti- allergic medicine, actinotherapy and chronic respiratory organ diseases are all taken into consideration. Call tree algorithm is used for the purpose of prediction that is illustrated in the figure below. The probabilities of patient can be predicted by the system that will give hope to the patients in future.

VIII. CONCLUSION

It can be concluded that, by using a method that integrates neural network classifier, binarization and GLCM, the accuracy of lung cancer detection process can be highly enhanced. This will subsequently help in the reduction of time required for cancer detection, ultimately making the system cost efficient. Further, if the patient is not detected with the lung cancer the system will proceed for the prediction process. Most state-of-the-art medical facilities are not available to patients in the remote regions but this proposed system would be web based. Therefore, even patients from remote areas with Internet connectivity can reap the benefits of this system thereby making the proposed system reach out to the wider public.

References:

[1]. Annangi, Pavan, Sheshadri Thiruvenkadam, Anand Raja, Hao Xu, XiWen Sun, and Ling Mao. "A region based active contour method for x-ray lung segmentation using prior shape and low level features." In 2010 IEEE international symposium on biomedical imaging: from nano to macro, pp. 892-895. IEEE, 2010.

[2]. Armato, Samuel G., Maryellen L. Giger, Catherine J. Moran, James T. Blackburn, Kunio Doi, and Heber MacMahon. "Computerized detection of pulmonary nodules on CT scans." Radiographics 19, no. 5 (1999): 1303-1311.

[3]. Besbes, Ahmed, and Nikos Paragios. "Landmark-based segmentation of lungs while handling partial correspondences using sparse graph-based priors." In Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, pp. 989-995. IEEE, 2011.

[4]. Chen, Hui, Yan Xu, Yujing Ma, and Binrong Ma. "Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on CT images: clinical evaluation." Academic radiology 17, no. 5 (2010): 595-602.

[5]. Cristobal, Gabriel, and Rafael Fonolla Navarro. "Space and frequency variant image enhancement based on a Gabor representation." Pattern recognition letters 15, no. 3 (1994): 273-277.

[6]. Dai, Shuangfeng, Ke Lu, Jiyang Dong, Yifei Zhang, and Yong Chen. "A novel approach of lung segmentation on chest CT images using graph cuts." Neurocomputing 168 (2015): 799-807.

[7]. El-Ba, Ayman, Georgy Gimel'farb, Robert Falk, Trevor Holland, and Teresa Shaffer. "A new stochastic framework for accurate lung segmentation." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 322-330. Springer, Berlin, Heidelberg, 2008.

[8]. Gao, Qixin, ShengJun Wang, Dazhe Zhao, and Jiren Liu. "Accurate lung segmentation for X-ray CT images." In Natural Computation, 2007. ICNC 2007. Third International Conference on, vol. 2, pp. 275-279. IEEE, 2007.

[9]. Henschke, Claudia I., David F. Yankelevitz, Ion Mateescu, Dean W. Brettle, Timothy G. Rainey, and Fred S. Weingard. "Neural networks for the analysis of small pulmonary nodules." Clinical imaging 21, no. 6 (1997): 390-399.

[10]. Hua, Panfang, Qi Song, Milan Sonka, Eric A. Hoffman, and Joseph M. Reinhardt. "Segmentation of pathological and diseased lung tissue in CT images using a graph-search algorithm." In Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, pp. 2072-2075. IEEE, 2011.

[11]. Itai, Yoshinori, Hyoungseop Kim, Seiji Ishikawa, Shigehiko Katsuragawa, Takayuki Ishida, Katsumi Nakamura, and Akiyoshi Yamamoto. "Automatic segmentation of lung areas based on SNAKES and extraction of abnormal areas." In Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference on, pp. 5-pp. IEEE, 2005.

[12]. John, Jibi, and M. G. Mini. "Multilevel Thresholding Based Segmentation and Feature Extraction for Pulmonary Nodule Detection." Procedia Technology 24 (2016): 957-963.

[13]. Kawata, Yoshiki, Noboru Niki, Hironobu Ohmatsu, Masahiko Kusumoto, Ryutaro Kakinuma, Kensaku Mori, Hiroyuki Nishiyama, Kenji Eguchi, Masahiro Kaneko, and Noriyuki Moriyama. "Computerized analysis of 3-D pulmonary nodule images in surrounding and internal structure feature spaces." In Image Processing, 2001. Proceedings. 2001 International Conference on, vol. 2, pp. 889-892. IEEE, 2001.

[14]. Kockelkorn, Thessa TJP, Eva M. van Rikxoort, Jan C. Grutters, and Bram van Ginneken. "Interactive lung segmentation in CT scans with severe abnormalities." In Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on, pp. 564-567. IEEE, 2010.

[15]. Lo, Shih-Chung Benedict, Li-Yueh Hsu, Matthew T. Freedman, Yuan Ming Fleming Lure, and Hui Zhao. "Classification of lung nodules in diagnostic CT: an approach based on 3D vascular features, nodule density distribution, and shape features." In Medical Imaging 2003: Image Processing, vol. 5032, pp. 183-190. International Society for Optics and Photonics, 2003.

[16]. Matsuki, Yuichi, Katsumi Nakamura, Hideyuki Watanabe, Takatoshi Aoki, Hajime Nakata, Shigehiko Katsuragawa, and Kunio Doi. "Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: evaluation with receiver operating characteristic analysis." American Journal of Roentgenology 178, no. 3 (2002): 657-663.

[17]. McNitt-Gray, Michael F., Eric M. Hart, Nathaniel Wyckoff, James W. Sayre, Jonathan G. Goldin, and Denise R. Aberle. "A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results." Medical physics 26, no. 6 (1999): 880-888.

[18]. Nakamura, Katsumi, Hiroyuki Yoshida, Roger Engelmann, Heber MacMahon, Shigehiko Katsuragawa, Takayuki Ishida, Kazuto Ashizawa, and Kunio Doi. "Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks." Radiology 214, no. 3 (2000): 823-830.

[19]. Pu, Jiantao, Justus Roos, A. Yi Chin, Sandy Napel, Geoffrey D. Rubin, and David S. Paik. "Adaptive border marching algorithm: automatic lung segmentation on chest CT images." Computerized Medical Imaging and Graphics 32, no. 6 (2008): 452-462.

[20]. Shi, Yonghong, Feihu Qi, Zhong Xue, Liya Chen, Kyoko Ito, Hidenori Matsuo, and Dinggang Shen. "Segmenting lung fields in serial chest radiographs using both population-based and patient-specific shape statistics." IEEE Transactions on Medical Imaging 27, no. 4 (2008): 481-494.

[21]. Silveira, Margarida, Jacinto Nascimento, and Jorge Marques. "Automatic segmentation of the lungs using robust level sets." In Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, pp. 4414-4417. IEEE, 2007.

[22]. Sofka, Michal, Jens Wetzl, Neil Birkbeck, Jingdan Zhang, Timo Kohlberger, Jens Kaftan, Jérôme Declerck, and S. Kevin Zhou. "Multi-stage learning for robust lung segmentation in challenging CT volumes." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 667-674. Springer, Berlin, Heidelberg, 2011.

[23]. Sun, Shanhui, Christian Bauer, and Reinhard Beichel. "Automated 3-D segmentation of lungs with lung cancer in CT data using a novel robust active shape model approach." IEEE transactions on medical imaging 31, no. 2 (2012): 449-460.

[24]. Suzuki, Kenji, Feng Li, Shusuke Sone, and Kunio Doi. "Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network." IEEE Transactions on Medical Imaging 24, no. 9 (2005): 1138-1150.

[25]. Yankelevitz, David F., Anthony P. Reeves, William J. Kostis, Binsheng Zhao, and Claudia I. Henschke. "Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation." radiology 217, no. 1 (2000): 251-256.

[26]. Yim, Yeny, Helen Hong, and Yeong Gil Shin. "Hybrid lung segmentation in chest CT images for computer-aided diagnosis." In Enterprise networking and Computing in Healthcare Industry, 2005. HEALTHCOM 2005. Proceedings of 7th International Workshop on, pp. 378-383. IEEE, 2005.