ICTIMES 2019

5th INTERNATIONAL CONFERENCE ON

Trends in Information, Management, Engineering and Sciences

Proceedings of International Conference on
Emerging Technologies in Computer Science (ICETCS)
- ISBN: 978-93-88808-61-3

Convenor
Dr. M. Narayanan
HOD-Dept. of CSE

Editor
Dr. J. Gladson Maria Britto
Professor - CSE

Editor in Chief
Dr. P. John Paul
Principal

Organized by

Estd :2005

MALLA REDDY COLLEGE OF ENGINEERING

Approved by AICTE - New Delhi, Affiliated to JNTUH, Accredited by NBA & Accredited by NAAC.
ISO 9001:2015 Certified Institution, Recognition of College under Section 2(f) & 12 (B) of the UGC Act,1956.

website: www.mrce.in

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

SMART WAY TO COMMUNICATE THE PASSENGERS IN THE TRAIN USING THE HYBRID TECHNOLOGY

Dr.T.Sunil
Sunil.tekale2010@gmail.com

U. Nagaiah
nagaiah1212@gmail.com

Abstract:

In order to cater to the needs of passengers travelling in the train by reserving the seats, the system is designed so as to provide passenger specific information as well as common information to all the passengers travelling. Passenger specific information can be like information in advance about the destination along with the information of the live train. Whereas the common information to the passengers can be like information about the live train. The system is designed such that it works on time every time because of the technology used which is called the hybrid technology. Here combination of internet and intranet technology is used. The passengers can avail this facility by opting the same during the seat reservation process.

Introduction:

Providing information to the passengers in the train is very important now days because of various reasons. Which is done and the facility is available to the passenger with the help of internet. But I want to provide the facility to the passenger who is travelling without making use of internet directly. The touch screen device will provide information related to location of the train and will also provide information about the destination of the passenger. The touch screen device when touched will provide this information. Apart from that it will also alert to the passenger about the destination in advance by means of alarm. As internet connectivity may not be available at all location especially in the forest area, this method of using intranet communication will provide the best solution for the same. This is an innovative method of providing information /alertness to the passenger without using the mobile device.

The basic information of passengers performing reservation will be available with the railway authority and this has to be exported to the main server arranged for a particular train. The server has to import the data and has to transfer the same to various clients which are located at every passenger seat along with the common client for each block of the bogie. The purpose of client common to block of each bogie can be referred as common client system is to provide general information common to all the passengers , whereas the purpose of the individual client is to provide particular information to specific client . The server will be updated every time from the main server and the same information will be used to provide the alerts and alarm to the specific passenger. The communication between the server in the train and the client can be done using intranet communication system and between two servers with internet. The basic idea which is novel is to see that the information which has to be provided to the passenger should be given on time and every time. The various devices used are touch screen devices with alarm facility (speakers) referred as clients along with the hubs at every bogie and the main server in the train. The communication can be with the help of

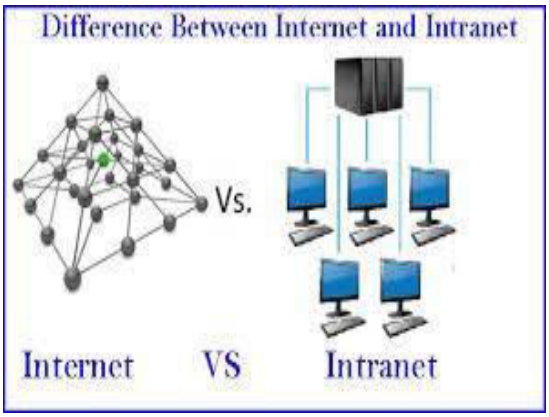
wires as well wireless. So as soon as the train is ready for departure at source station the information will be imported and the same can happen at multiple other stations. The passenger will be given an option to select the facility while the passenger performs the reservation. This facility will be provide only to the passengers travelling by purchasing tickets in advance (reserved). Where as in general compartments the arrangement can be done only for the common clients which will inform about the position of the live train.

Design consideration:

1. The system is designed with the help of internet and intranet technology to provide passenger specific and common information.
   a. As the chances of designing with internet technology may fail in many locations.
   b. The system is designed with touch screen devices so as to make it easier to use for the common man or for a passenger who is not well versed with technology.
   c. The system will provide different information to the clients according to the destination chosen.
   d. The System designed helps for specific individual as well as common to all.
   e. Usage of hybrid technology is novel so as to see that the system works on time every time.
   f. The system is easily managed because of having total control of data.
2. The system is designed using intranet so that the same works in all the geographical locations where internet do not /may not work.
3. A system as said in claim 1 is designed by using touch screen device is an innovative idea/novel idea as the common man faces problems with the latest technology.
4. The system designed is very innovative in terms of providing information as the destination chosen by the passenger. The system will alert and guide the passenger in advance, so that there will not a problem at the last minute.
5. The designer of the system is very complicated and innovative as the same system has to provide information to individual passengers when prompted as well as to all passengers sitting in the compartment at specific time interval.
6. Designing the system with hybrid technology was innovative concept so that we can have better and concrete system at place.
7. The algorithm used here will help the build a robust system.
8. The system is designed so as to cater to the need of people who are not well versed with technology.

Fig.1 Image of Intranet used in the process of passenger communication

Fig.2 Image of internet and intranet used in the process of passenger communication



Challenges:

To store the data from the source station before the train starts and to feed the same in the main server which will be placed in the drivers cabin and to see that the cables laid out for passenger communication are not disturbed or cut by any means. The touch screen device should be made very tough or robust so that it will be damaged easily by the passenger wanted or by mistake.

Existing System:

In the existing system the passengers are made to depend on the mobile phones or they need to depend on another passenger to provide information regarding the current location of the train as well as to wake up in the morning for the destination station. The problem here is if the mobile network is not working than there is every chance of system being collapsed. So in order to reduce the dependency the new system will help to provide information to the passenger as and when they required as well as they can get the information half an hour before they actually have to get down the station.

How the system works:

As the system uses intranet and internet technology which otherwise can be called as hybrid technology it is very easy to provide information to the passenger in much better way. The information will be collected at the source station with the help of internet facility and the information will be shared among the various passengers with the help of intranet facility along with the touch screen devices arranged.

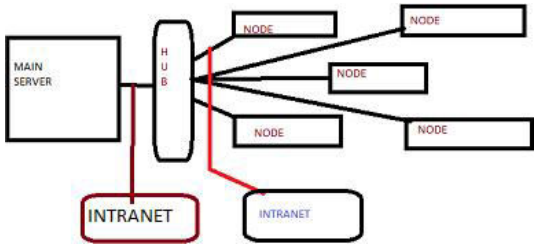
The information which is the data pertaining to the passengers reserved for that particular train will be gather and stored in the server system at the source station, then all the nodes which are connected to the hubs in various bogies are connected. The nodes will get the information from the server using intranet technology. So here the data is stored in the server with the help of internet and the same is shared among various nodes using intranet technology, then the dependency on the mobile technology and mobile network is completely zero.

Fig: 3 shows the touch screen which will be used by the passenger.

Results:

The data is stored in the server system from the main railway reservation system using internet technology which can be referred as MD1 and then using the intranet technology between the server and the various hubs and the nodes connected to the hubs the data is transferred from the main server to the nodes via hubs which can be referred as ND1…NDn.

Fig: 4 Refers to communication between the main server and the nodes present in various bogies.



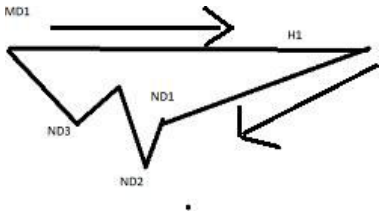
MD1-> ND1 MD1->ND3

MD1->ND2 MD1->ND4

So the information is passed from the main server i.e MD1 to rest of the nodes including the hubs and accordingly the passenger can get the information regarding the various stations and the distance from the destination.

MD1>H1>ND1 MD1->H1->ND2

Fig:5 How the data flows from MD1 to H1 and from there ND1…NDn



Information between the MD1 and the H1 which is the hub is communicated using intranet and from there to the various nodes is also by intranet technology only.

Whereas the information from the main reservation server to the main server in the train is communicated using internet technology.

Conclusion:

When the system is designed it will be very useful for the passengers who are travelling by taking the upper part of the berth because there is no other way for them to get the information related to the places they are visiting and arriving. This particular system will help them to get information regarding the position of the train and also can get the information about their destination in advance. The passenger for this may have to shell minimum amount which can be used for the development of the system in future. Overall the conclusion is that this is going to be well defined asset for the passengers as well as the railway board as it helps to increase the revenue and passengers will be happy because o getting right information at right time. The passengers will be provided with touch screen for this purpose.

REFERENCES

[1] Carol L. Schweiger, "Real-Time Bus Arrival Information Systems - A Synthesis of Transit Practice",

Transportation Research Board, 2003.

[2] "Review of Current Passenger Information Systems", Prepared for the INFOPOLIS 2 Project (No. TR 4016),
Deliverable 1, WP03, Info polis 2
Consortium, August 1998.

[3] Hu, K. and C.K. Wong, "Deploying Real- Time Bus Arrival Information and Transit Management Systems in Los Angeles", abstract prepared for the ITS America
12th Annual Meeting, Long Beach, Calif., April 29-May2, 2002.

[4] Helsinki City Transport System http://www.hel2.fi/ksv/entire/repPassenge rInformation.htm

[5] Tolarno Inc. - Passenger Information Services http://www.telargo.com/solutions/passene r_informationservices.aspx

[6] Terron Microsystems Pvt. Ltd.- GPS based Passenger Information System for Buses http://www.terronmicrosystems.com/prod ucts.php

[7] Brendan Kidwell, "Predicting Transit Vehicle Arrival Times", GeoGraphics Laboratory, Bridgewater State College, August 2001.

[8] Lin, W.H. and Zeng, J., "Experimental Study of Real-Time Bus Arrival Time Prediction with GPS Data."

Transportation Research Record, 1666, pp. 101-109, 1999.

[9] Chien, I. J. S. and Ding Y., "Applications of Artificial Neural Networks in Prediction of Transit Arrival Times", 1999 Annual Meeting of ITS America, Washington D. C., 1999.

[10] Wanli Min et al, "Road Traffic Prediction with Spatio-Temporal Correlations", IBM Watson Research Center, 2007.

[11]A. Guin, "Travel Time Prediction Using a Seasonal Autoregressive Integrated Moving Average Time

Series Model", IEEE ITSC, 2006.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

An Efficient method to Transfer Files in Peer- to-Peer Networks over Video on Demand Services

[1]M.Narayanan, [2]E.Lingappa
[1]Professor, [2]Assistant Professor
Department of Computer Science and Engineering,
Malla Reddy College of Engineering, Maisammaguda,
Dhulapally, post via Kompally, Secunderabad, Telangana

Abstract

One of the most promising technology is Peer-to-Peer. Each peer acts as a client and server so every peer is accounted for the downloading and uploading of files. Though frequently compared to client server model it is altogether distinct to it. Therefore it is also known as single system program so at that instance each peer can act as client/server. Content distribution or file sharing or dissemination of information of the file is one of the most important application. Besides this, the other applications include publications and dissemination of software, content delivery network, streaming media and multicast streaming. It also allows on demand content delivery. Science networking, searching and communication are other applications of peer to peer.

Keywords: P2P Network, Client server Model, Hybrid Model, Searching Communication

I. INTRODUCTION

The concept of Peer-to-Peer file sharing technology is growing at an unprecedented rate. Each peer can perform like both client and server and that sets them apart from the client server model. File can be shared depending upon the bandwidth and the peer possessing larger bandwidth is given the precedence followed by the other. Suppose the key server is sharing the file to client and another peer requests for file sharing, the key server will put together an arrangement for the requested peer. And the client receiving the file at this time will transfer the file to the client which is making the request. In this scenario it is free for anyone to operate as a client and server. Therefore the amount of the file transferred will be larger and at a higher speed. It comes handy for downloading a video file from the net. The files can be shared by utilizing the techniques such as MCC, FIFO, LFU, and LRU. When a file wishes to share from the server, it will apparently look for the peer which possess a higher bandwidth [15].

Peers with similar bandwidth experiences LFU (low frequently used). So the peer which is not used frequently will receive the file. In peer to peer the files are stored in the cache memory in the server. When the cache memory is completely occupied it will then erase the files on the basis of first in first out (FIFO). The file which is deleted first will be the one which was first in. It is the responsibility of the cache to locate the client having the necessary file. This memory enables the files to be transferred between peer to peer. The capacity of the cache memory is up to 1024mb equivalent to 1kb.It will follow the following technique when the cache is completely occupied. Peer to peer enables many users to work simultaneously. Therefore it is considered as the best file sharing network. Depending on the bandwidth, the peers will be separated as clusters [2] with each cluster consisting a definite bandwidth.

For example cluster1 contains 1mbps bandwidth, cluster 2 contains 2mbps and similarly each clusters comprises of different bandwidth. Within each cluster many peers will be separated wherein every peer will possess similar bandwidth. In this peer to peer, all the peers will be linked in corresponding manner [1], [8]. There is no restriction as who should act as a client and who should be a server. This property enables to not only share video file but even some digital files and some computer files, books, movies, music and games. This particular P2P software enables to look for client at nearest proximity in the P2P network and conveys the file. The Fig.1 shows the client Server Model and peers (nodes) of P2P network will be end user computer system interconnected between the internets. Scientist Gnutella and kazaa were the first to develop this new P2P file sharing method.

ISBN 978-93-88808-61-3

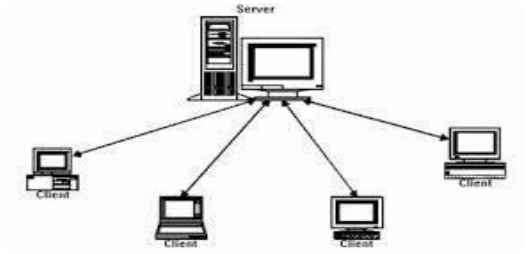


Fig 1: Client Server Model

II. RELATED WORK

In the year 1999, the earlier peer to peer file sharing method was developed by NAPSTER. Following the death of NAPSTER, gnutella and kazaa developed the new peer to peer file sharing network [7] . The old Peer to Peer file sharing was based on the central server system and enabled sharing of only music files effectively. Following extensive research, the system has developed to a great extent and any kind of files can be transferred utilizing this peer to peer file sharing technique. Unlike the client server model, this method is connected parallel and is known as decentralized and dispersed design [3][1].

A. Mentioned below are the two types in P2P

The structured p2p is organized by some definite standard and some algorithms. Some of the real time examples of structured P2P are some of the network computer workstations [5]. Unstructured P2P computer architecture consists of three types of models as follows: It is a democratization of every peer group nodes. They are two forms in order to accomplish the route:

One of the probable structuring is direct messaging. It is communicated through peer group members until an object of the member group is found. Further it must establish the members in HORIZON group. Horizon means limit of visibility from the node generating the query.

Another probable way for attaining the routing structure is distributed catalog. It necessities an energetically balanced catalog because it is indexed as parameter and searches as distributed catalog. This may also be not very efficient but it is very safe to work.

It is very necesary to enhance the P2P models in order to improve the search potential and system performance. Data accumulation in Peer-to-Peer is enormous. So in order to address these issues, this paper presents the application of data mining technology to P2P network. Based on SWLDRM, few developments are considered and nodes are clustered depending on the characters of object stored by K-NN algorithm. Simulation result proves that K-chord is more efficient in terms of performance and search when compared to SWLDRM [9].

One of the extensive classification of community based P2P systems is presented in this paper. Users belonging to a definite network forms this community. And the augmentation of these communities is influenced by factors such as value of the content, projection for enhanced performance and user experience enhancement. A campus network and a national ISP located in diverse continents are the two distinct environments that this study focuses. Here, the key P2P systems are found to be large scale closed communities. Results confirm that traffic on Internet peering links are reduced by localizing traffic inside ISP boundaries [10].

It is becoming increasingly challenging for ISPs because of the unprecedented growth of P2P applications consuming humongous bandwidth resources. So, it is paramount to handle P2P traffic efficiently and simultaneously protecting the P2P user interest. This paper primarily focus on analysing current mainstream P2P optimization strategies and implement NRDA technology. This technology permits ISPs to manage P2P traffic efficiently and independently on certain links and also to respond quickly in case of an abrupt necessity of resources. By and large, the planning of network resources are executed by the NRDA and it also offers resource planning capability to the ISPs [11].

In today's internet world, ISPs are facing an uphill task of providing basic network services for P2P users and also to effectively manage network bandwidth usage. But, existing strategies fail to fulfil these requirements. This paper proposes to devise a plain and efficient system for ISPs to strike a balance between service and network management. This can be attained by suggesting a file-aware P2Ptraffic classification method which can identify files and the associated flows. Two alternatives are proposed. One is by limiting the per-file bandwidth consumption and the other by measuring a real-life trace from peers and files perspective. The results show that as per the actual demands, ISPs can expediently choose suitable traffic management parameters [12].

Most of the video streaming services are developed for wired networks. But the challenge is to stream it in wireless environment which demands several alterations. Through an logical representation, this paper proposes a performance evaluation model of the traffic behaviour which bears a resemblance to the network interactions during a video transmission. In wireless environment, when the number of nodes increases the quality of video degrades due to collisions. Therefore, *P2P-TV* applications should be integrated from lower layers to meet the level of quality requirements. The throughput parameter is determined by this model and the network performance is evaluated precisely [13].

There should be monitoring of the Service Level Objectives in order to meet the Service Level

Agreements (SLAs) and regulate vital network services. However, the probing techniques are expensive and are also labour-intensive and prone to error. So, Peer-to-Peer (P2P) technology is employed to improve the detection of SLA breach. A P2P management overlay is considered to coordinate the probe activation and share measurement results between the network devices. In large scale networks, an autonomic P2P solution is proposed to coordinate active measurement probes. The solution is proved to be feasible as per the simulation results [14].

B. Centralized P2P

It consists of a central look up server linked in a star network style. In this type of peer to peer model, the message can be sent with ease and with greater speed because of less traffic. Because of the fact that it has less traffic it can be quickly addressed. But bottleneck behaviour which is a single point of failure is the biggest drawback. Though addressing is very efficient it is not safe to work. The Fig.2 shows the centralized P2P [4].

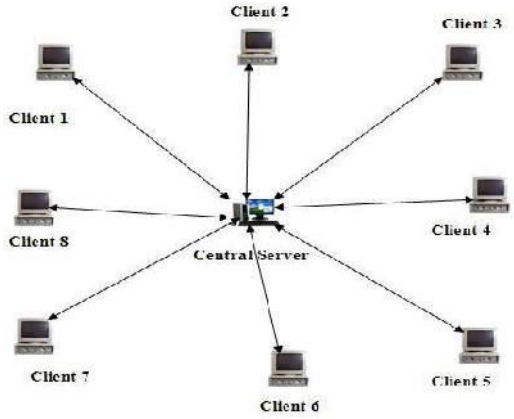


Fig.2: Centralized P2P

C. Hybrid P2P

It is an overlay routing structure consisting of super peers and leaf nodes. Hence leaf nodes are connected in star network and super peers act as shield to the leaf nodes [8]. Addressing is partly efficient and it is not safe.

Furnished below are few of the benefits of the peer to peer network:

a. Operating system is not needed in this network.
b. Unlike the client server model, a specialized complex network set up is not required. It can be created with ease and does not demand any superior knowledge.
c. Server is not so pricey because of the fact that an individual client can act as a server when uploading a file to another user.
d. If one of the peers does not succeed in sharing, it will not disturb any other part of the network. It would be just occupied with another work of sharing and will remain unavailable to other peers.
e. Compared to the client server model, the file sharing will be at a higher speed.
f. Easy availability and reduced cost is another big benefit of peer to peer network. So it enables users to utilize it in low cost.
g. Peer to peer has well tested peers, it will not have the peers which is not ready to share the file among another peer i.e. well tested simplicity.
h. It doesn't require a dedicated server. Anyone can act as server and anyone can act as client, so any computer can access both server and workstation.

D. Problems of P2P networking

Some of the potential problems faced by the users with Peer to Peer software are Bandwidth utilization, copyright infringement, and security issues. It also face some troubleshooting network problem, A computer may fail to process for many reasons this one of the basic reason for the problems in P2p networking. If it is not working problem it will affect the entire home networking to stop functioning. The Fig.3 illustrated the model diagram of P2P Networks.

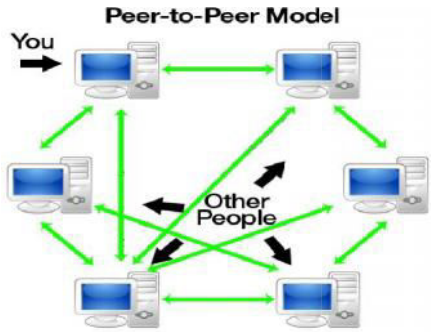


Fig.3: Model diagram of P2P networks

III. PROPOSED MODEL FOR P2P SYSTEMS

A. Core Transfer Engine Layer

The requested files between peers is transferred by this layer and it also carry out all the responsibilities of Peer actions. In this particular project, we anticipate to see some codes around Peers. This part is the heart of this system. When a peer commences its work, the first task is to register itself as a peer. Subsequently it should play the part of both server and client [6] Later, if any peer asks a file, first it should search the file and after receiving the file's information such as the destination peer host name, it should utilise that information to connect to the peers and then download the file. The PNRP Manager (Peer Name Resolution Protocol) class is responsible to Register and Resolve peers.

ISBN 978-93-88808-61-3

The Register () method registers the peer in the cloud and accepts a list of Peer Info type as its input argument.

B. Proposed Algorithm: Algorithm for Peer Registration

```
Algorithm_PeerInfo_Registration
Begin
List PeerInfo, RegisteredPeer;
ForEach Registration;
String TimeStamp;
String TimeStamp = String.Format("FreeFile
Peer Created at :Zeroth Position")
DateTime.Now.ToShortTimeString ();
Registration.Comment = TimeStamp;
Try
Registration.Start();
Begin
IF RegisterdPeer.FirstOrDefault
X is HostName is
        Equal to Registration_PeerName
        And PeerHostNamePeerInfoPeerInfo =
        New PeerInfo
(registration.PeerName.PeerHostName,
Registration.PeerName.Classifier,
Registration.Port);
PeerInfo.Comment equal to
Registration.Comment;
RegisterdPeer.Add(peerInfo);
EndIF
 End
EndFor
End
```

C. File Transferring service in P2P systems

In order to give the essential files to other peers, the File Transfer Service Host class enables each peer as a server host. The TCP protocol is utilized for conveying the data among peers. Depending on the peer host name the DoHost() method gets an address. Subsequently an interface is added who applied the Service Contract feature. Hence, to make its methods reachable around service each peer publishes a service to the outside world. The Table 1 illustrated algorithm for file transferring service in P2P systems

TABLE 1: ALGORITHM FOR FILE TRANSFERRING SERVICE IN P2P SYSTEMS

```
Sealed class FileTransferServiceHost
{
   Public void DoHost(List<PeerInfo> peers)
   {
      Uri [] Uris = new Uri[peers.Count];

      String Address = string.Empty;
      For (inti = 0; i<peers.Count; i++)
      {
         Address = string.Format("net.tcp://{0}:{1
}/TransferEngine",
peers[i].HostName, peers[i].Port);
         Uris[i] = new Uri(Address);
      }

FileTransferServiceClass
currentPeerServiceProxy = new
FileTransferServiceClass();
ServiceHost _serviceHost = new
ServiceHost(currentPeerServiceProxy, Uris);
NetTcpBindingtcpBinding = new
          NetTcpBinding(SecurityMode.None);     }
}
interfaceIFileTransferService
{
   [OperationContractAttribute(IsOneWay = false)]
byte[] TransferFileByHash(string fileName,string
hash, long partNumber);

   [OperationContractAttribute(IsOneWay = false)]
byte[] TransferFile(string fileName, long
partNumber);
}
```

IV CONCLUSION

One of the most promising technology in the internet world is the P2P network. After the multi-core processor is developed the P2P network will achieve unprecedented worth in the web. This study is a complete survey paper providing details about P2P networks. My survey provides types of P2P computing algorithms. Currently every field depends on computer and its various application. Therefore it is highly essential to design state-of-art P2P systems for video on demand service.

REFERENCES

[1]. Biazzini, M.; INRIA-Bretagne Atlantique, France; Serrano-Alvarado, P.; Carvajal-Gomez, R.,"Towards improving user satisfaction in decentralized P2P networks", in Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference , 20-23 Oct. 2013.

[2]. Garant, D. ; Dept. of Comput. Sci., USNH, Keene, NH, USA; Wei Lu,"Mining Botnet Behaviors on the Large-Scale Web Application Community ", in Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference ,25-28 March 2013.

[3]. Hirave, T.; Dept. of Comput. Eng., Mumbai Univ., Mumbai, India ; Surve, S. ; Malgaonkar, S.,"Selecting efficient peers in P2P networks for parallel task computing",in Advances in Technology and Engineering (ICATE), 2013 International Conference ,23-25 Jan. 2013.

[4]. Biazzini, M. ; INRIA-Bretagne Atlantique, France ; Serrano-Alvarado, P. ; Carvajal-Gomez, R.,"Towards improving user satisfaction in decentralized P2P networks", in Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference , 20-23 Oct. 2013.

[5]. Garant, D. ; Dept. of Comput. Sci., USNH, Keene, NH, USA ; Wei Lu,"Mining Botnet Behaviors on the Large-Scale Web Application Community ", in Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference ,25-28 March 2013.

[6]. Hirave, T. ; Dept. of Comput. Eng., Mumbai Univ., Mumbai, India ; Surve, S. ; Malgaonkar, S.,"Selecting efficient peers in P2P networks for parallel task computing",in Advances in Technology and Engineering (ICATE), 2013 International Conference ,23-25 Jan. 2013.

[7]. Ripeanu, M. ; Dept. of Comput. Sci., Chicago Univ., IL, USA,,"Peer-to-peer architecture case study: Gnutella network", in Peer-to-Peer Computing, 2001. Proceedings. First International Conference , 27 Aug 2001-29 Aug 2001.

[8]. Kang Chen ; Dept. of Electr. &Comput. Eng., Clemson Univ., Clemson, SC, USA ; HaiyingShen ; Haibo Zhang, "Leveraging Social Networks for P2P Content-Based File Sharing in Disconnected MANETs" , Mobile Computing, IEEE Transactions on (Volume:13 , Issue: 2), Date of Publication : 26 November 2012

[9]. Zhu, Xiaoshu, Miao Xie, and Tang Lu. "Application Research of Data Mining Techniques in P2P Network." In *Genetic and Evolutionary Computing, 2009. WGEC'09. 3rd International Conference on*, pp. 297-300. IEEE, 2009.

[10]. Torres, Ruben, Marco Mellia, Maurizio M. Munafo, and Sanjay G. Rao. "Characterization of community based-P2P systems and implications for traffic localization." *Peer-to-Peer Networking and Applications* 6, no. 2 (2013): 118-133.

[11]. Ma, Dongchao, Xiaoliang Wang, Wenlong Chen, Shen Yang, and Li Ma. "A research on dynamic allocation of network resources based on P2P traffic planning." *Peer-to-Peer Networking and Applications* 7, no. 4 (2014): 511-524.

[12]. Song, Tian, and Zhou Zhou. "File-aware P2P traffic classification: An aid to network management." *Peer-to-Peer Networking and Applications* 6, no. 3 (2013): 325-339.

[13]. Urrea Duque, Juan Pablo, and Natalia Gaviria Gomez. "Throughput analysis of P2P video streaming on single-hop wireless networks." In *Communications (LATINCOM), 2014 IEEE Latin-America Conference on*, pp. 1-6. IEEE, 2014.

[14]. Nobre, Jéferson C., Lisandro Z. Granville, Alexander Clemm, and Alberto Gonzalez Prieto. "On the Use of Traffic Information to Improve the Coordinated P2P Detection of SLA Violations." In *Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on*, pp. 613-620. IEEE, 2014.

[15]. M.Narayanan, C.Arun" An Efficient Method for Handling Data Segment with Multi-Level Caching over Video-on-Demand using P2P Computing", *European Journal of Scientific Research*. ISSN 1450-216X Vol. 93 No 2 December, 2012, pp.206-213

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

IOT BASED SMART AGRICULTURE MONITORING AND COMTROL SYSTEM

[1]Dr.J.GladsonMariaBritto,[2]Dr.Bhoopathy.V,
[1]Professor,CSE, Malla Reddy College of Engineering
[2]Professor,CSE, Malla Reddy College of Engineering

Abstract: The Major developing advancement technology in upcoming future is Internet of Things, commonly known as IOT is a promising area in technology that is growing day by day. Agriculture plays vital role in the development of agricultural country. In India about 80% of population depends up on farming and one third of the nation's capital comes from farming. The problems based on agriculture have been always hindering the development of the country. The highlighting features of this project includes wireless network sensors to connect multiple sensors data and to display big-data through Thing speak channel software to perform tasks like weeding, spraying, moisture sensing, bird and animal scaring, keeping vigilance. The development includes smart irrigation with smart control and intelligent decision making based on accurate real time field data. Which also includes temperature maintenance, humidity maintenance and weather reports. Controlling of all these operations will be through any smart mobile or computer connected to Internet and the operations will be performed by interfacing sensors. The data can be completely updated faster when compared to other wireless computing.

Keywords: *Internet of Things, Wireless sensor Networks, Micro keil version, Thing speak, Big data collection, cloud computing.*

INTRODUCTION

Agriculture is the unquestionably major process provider in India. With rising population, there is a need for increased agricultural production. In order to support greater production in farms, the requirement of the amount of fresh water used in irrigation also rises. Currently, agriculture accounts 93% of the total water consumption in India. Unplanned use of water continuously results in wastage of water. This suggests that there is an urgent need to develop systems that prevent water wastage without imposing pressure on farmers. 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. Farmers could be able to smear the right amount of water at the right time by irrigation. Avoiding irrigation at the wrong time of day, reduce run off from overwatering saturated soils which will improve crop performance. The available traditional methods of irrigation are drip irrigation, ditch irrigation, sprinkler system. This problem can be easily rectified by making use of the automated system rather than the traditional systems. The current

irrigation methodology adopted employ uniform water distribution which is not optimal. The client and server programming condition which could improve the data over big data collection` Large number of entries could be over seen with online portal services. In addition to the standalone monitoring station, Wireless Sensor based monitoring system been developed which is composed of number of wireless sensor nodes and a gateway. This system here provides a unique, wireless and easy solution with better spatial and temporal resolutions. This could also include the farm security from animals attack without injuring the animals like in manual method.

Motion detector is used to sense the temperature of the animals and to be thrown away from farm land. The farmer is notified about the decision whether to irrigate or not through a either a web app or mobile app which is developed using WEB. Based on the decision received from the machine learning process, The farmer can trigger the irrigation process through his mobile phone .same is also provided through a web interface.

LITERATURE SURVEY

The older method and one of the oldest ways in agriculture is the manual method of checking the parameters. In this method the farmers only by themselves verify all the parameters and calculate the readings that's why to overcome this stress and relief from stress, It focuses on developing devices and tools to manage, display and alert the users using the advantages of a wireless sensor network method. It aims at making agriculture smart and modern using automation and IoT technologies. It provides a low cost and effective wireless sensor network technique to acquire the soil moisture and temperature from various location of farm and as per the need of crop controller to take the decision whether the irrigation is enabled or not. It proposes an idea about how automated irrigation system was developed to optimize water use for agricultural crops. In addition, a gateway unit handles sensor information. The atmospheric conditions are monitored and controlled online by using Ethernet IEEE 802.3.It is designed for IoT based monitoring system to analyze crop environment and the method to improve the efficiency of decision making by analyzing harvest statistics. The source of power 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 .Various techniques agricultural applications like seed sowing, sloughing, water irrigation, crop cutting and etc. like this several operations were done with IOT. Various companies in INDIA and globally have been proposed in using micro controller based controllers for various have come with novel solutions using automated systems for various application with specific individually (www.smartagriculture.com). Most work carried out in literature and organizations have their inherent advantages and disadvantages. These manufacturers do not have multiple agricultural applications integrated in a single hardware. To eradicate such errors or disadvantages we are introducing a multi functional design using wireless sensor networks. The system was based upon an automated irrigation system by using mainly a soil moisture sensor and an Android smart phone. With this system, people can have a better control on their irrigation time and can also save water. In this prototype, different soil samples and crops for calibration at various moisture levels was tested. However, to improve this analysis, various soil samples from different places could have been tested and also during different weather conditions. Apart from soil moisture, other factors of the soil could have also been monitored. The cloud computing that could improve the advanced technologies using big data

ISBN 978-93-88808-61-3

collection through the fast updates of data's through online entries. In an updated wireless network sensor system, the data that could be updated through faster applications.

PROPOSED SYSTEM

The system is a combination of hardware and software components. The hardware part consists of programming microcontroller AT-mega328 which could connect the other sensors to collect the data. Solar panel which act as an renewable source of energy that is to be connected as an rechargeable power source to save the power. The moisture sensor that could recover the dry or wet condition of the soil and thereby the intimation could be sent to the farmers through web browser or GSM module. Then the centrifugal pump could be turned ON/OFF by the farmers from anywhere or anyplace through the online channel creation using Thing speak. Thing speak is the webpage designed using PHP. The webpage is hosted online and consists of a database in which readings from sensors are inserted using the hardware. soil moisture sensors along with LM38 comparator modules were placed in different soil conditions for analysis. It reads the moisture content around it. A current is passed across the electrodes through the soil and the resistance to is made the current in the soil determines the soil moisture. If the soil has more water resistance will be low and thus more current will pass through. An Intelligent IOT Based Automated Agriculture has been proposed so as to reduce the wastage of water and security to the crops. The system mainly monitors the behavior of soil moisture, air humidity, air temperature and secures the crops from animal attack and sees how it contributes to evaluate the needs of water in a plant. The data is taken from the sensors and is transferred through internet to the mobile application or web app and water pump turn ON/OFF using web app.

BLOCK DIAGRAM:

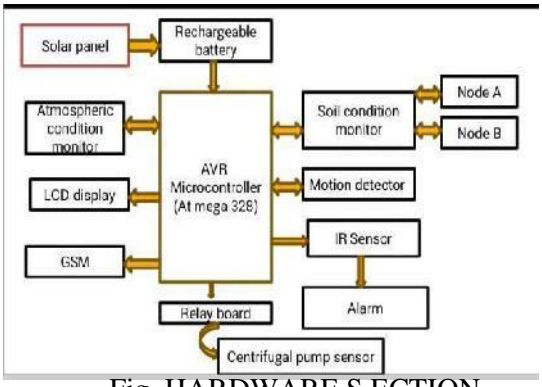


Fig. HARDWARE S ECTION

Which also includes temperature maintenance, humidity maintenance and weather reports. Controlling of all these operations will be through any smart phone will be performed by interfacing sensors. The WI-FI module that could be interconnected with the moisture sensor at various nodes of node A and node B. This collects the data continuously and to be entered through online portal using thing speak software. The major advantage of this method is secured and maintained complete data for farmers convenience.

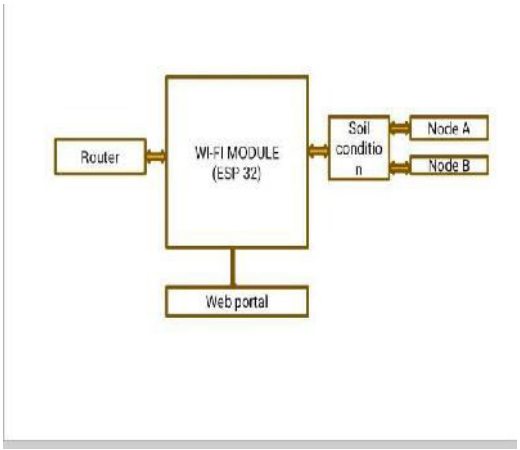


Fig , Block diagram of WI-FI section

ISBN 978-93-88808-61-3

Fig. ANIMAL ATTACK

MOTION DETECTOR : It will detect the thermal heat from animals body .GSM communication can be used to send immediate notification to farmer Alarm sounds can be activated.

DESCRIPTION: The AT-mega 328 is a low-power, high performance CMOS 28-bit microcontroller with 8K bytes of in-system programmable Flash memory. The device is manufactured using Atmel's high-density nonvolatile memory technology and is compatible with industry standard 80C51 instruction set and pin out. The Flash over on-chip allows the program memory to be reprogrammed in-system or by conventional non-volatile memory programmer. By combining a versatile 28-bit CPU with in-system programmable flash on a monolithic chip, the Atmel 328 is a powerful microcontroller which provides a highly flexible and cost effective solution to many embedded control applications. The analog to digital converter with 10 outputs has been connected to microcontroller which converts the analog data into digital format. The GSM sim 800 module along with WI-FI interconnect module that could store the data and send the information to the login channel to the farmers. In case of failure of the network connectivity immediately the GSM performs its operation by sending message to the farmers registered number. The combination of both i from the farm land internet and GSM could be performed based on its access and design over the data collection from the farm land.

Fig., GS M module to network connection

FUNTIONAL DESCRIPTION

WATER IRRIGATION

Water irrigation done through the basis of required to the plants and without wastage of the water. Such scientific method of water irrigation done by considering various parameters like soil type, crop type etc. The prevention of soil erosion practices which can drastically decrease negative effects associated with soil erosion such as reduced crop productivity, worsened water quality, lower effective reservoir water levels, flooding, and habitat destruction. Contour farming is considered an active form of sustainable agriculture.

SOFTWARE TOOLS

Software tools used

The software's which are used to developed this project are

- MicroKeil IDE compiler
- Languages used: Embedded C
- Things peak online web entry.

SQL Database and Power BI

At this point, the data found in the database needs to be transformed into a more user

represented as will not be understand SQL queries. Hence, to cater for this problem, Power BI is used to reconstruct the data into a visual representation such as a graph.

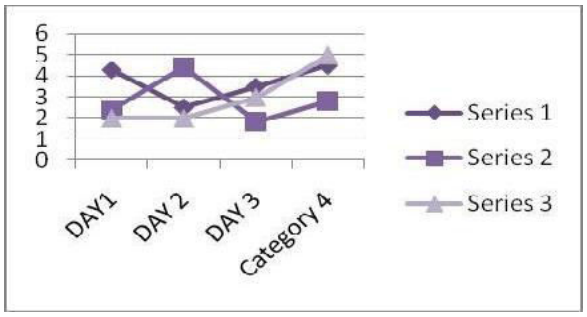


Fig. Graph Representation

Machine Learning

The Thing speak machine learning is the core logic of the proposed system. In general, a dataset is needed to train the machine to in the data in order to decide whether or not. For better precision, aOpenweathermap.com API is with the aim of knowing when the water pump needs to be opened. The pseudo code gives a simple illustration on how the machine learning system works producing code that is portable across wide platforms.

CONCLUSION

This multipurpose system gives an advance method to The system mainly monitors the behavior of soil moisture, air humidity, and air temperature and see how it contributes to evaluate the needs of water in a plant. The system uses machine learning and compares actual values obtained from sensors with a threshold value that has been fed to the machine learning for analysis. Next to this process, the machine learning cross checks the result obtained with weather forecast and then decides whether irrigation needs to be done or not. The farmer receives a notification on his smart phone and he can choose to turn on the water pump with a button click. Moreover, the system has a web app and is helpful if ever the farmer wants to see the statistical sensor data and assess the change in sensor readings throughout a time period. Furthermore, the system can calibrated for different type of plants, that is, the user is provided with a list of plants choices in his web app and mobile app. With this the user can choose the specific type of plant that is being cultivated and obtained threshold value and thus a more accurate irrigation prediction. Besides, an SMS system can be integrated if in case there is no internet connection. With this, the user would be notify about the prediction via an SMS and he can choose to switch on or off the water pump by replying to the SMS received. The entries that could also been saved safely for an farmers acknowledgement with date an time condition. The future works of transferring data is in the mode of social networks also through online data storage of cloud computation.

REFERENCES

1. *P. Narayut, P. Sasimanee, C.-I. Anupong, P. Phond and A. Khajonpong, 2016. A Control System in an Intelligent Farming by using Arduino Technology. Student Project Conference (ICT-ISPC), 2016 Fifth ICT International, pp. 53-56, 2016.*
2. *K. Benahmed, A. Douli, A. Bouzekri, M. Chabane and T. Benahmed, 2015. Smart Irrigation Using Internet of Things. Fourth International Conference on Future Generation Communication Technology (FGCT), 2015.*
3. *A.N. and K. D, 2016. Expe rimental investigation of remote control via Android smart phone of arduino based automated irrigation system using moisture sensor. 3rd International Conference on Electrical Energy Systems (ICEES), 2016.*

4. *T. Baranwal, N. and P. K. Pateriya, 2016. Development of IoT based Smart Security an d Monitoring Devices for Agriculture. 6th International Conference -Cloud System and Big Data Engineering (Confluence), 2016.*

5. *G. M.K., J. J. and A. M. G.S, 2015. Providing Smart Agriculture Solutions to Farmers for better yielding using IoT. IEEE Technological Innovation in ICT for Agriculture and Rural Development (TIAR), 2015.*

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

A SURVEY -EVENT DETECTION USING MACHINE LEARNING APPROACH IN CYBER-PHYSICAL SYSTEMS

[1]Dr.A.Kannagi
Professor, Department of CSE,
Malla Reddy College of Engineering,
kannagianbu_cse@mrce.in

[2]Mr.K.Praveen Reddy,
Assistant Professor, Department of CSE,
Malla Reddy College of Engineering,
praveenreddy_cse@mrce.in

Abstract

Asystematicapproach of extractingknowledgefromsensor data at various platforms plays a major role in the Data Mining Community to determine the intrusion detection in Cyber-Physical Systems (CPS), e.g., Assuming that there may be at whatever harm for building or aviation vehicles, those harm will be distinguished starting with the non stop arriving data. In the Existing methodology exhaustive Data Mining Framework which uses Differential Sensor Pattern (DSP) for Intrusion detection, DP miner has been used which greatly reduce the energy for calculation and correspondence in the CPS, the different pattern of sensors is been extracted that may have event information with a low communication cost but it can validate actual data only on lower data analysis whereas, for big data it cannot be sensed accurately. In order to achieve accuracy in big data environments, differential sensor mining technique with a machine-learning approach is been proposed for handling continuous quality improvement in event detection and it will useful for many CPS applications. .

KEYWORDS : Cyber-Physical Systems (CPS), Data Mining, Event Detection..

Introduction

Cyber-physical frameworks (CPSs) mix those learning What's more innovations of the third wave from claiming data processing, correspondence What's more registering with the learning Furthermore innovations of physical artifacts also engineered frameworks [1]. There appears with make a concurrence in the written works on the reality that those calling what's more learning of cyber physical frameworks would not monodisciplinary. However, it is at present debated if this discipline may be interdisciplinary, multi-disciplinary, and alternately trans-disciplinary to nature. Backers of the inter-disciplinary perspective contend that those mission fromclaiming CPSsscience Furthermore engineering organization is with make An span the middle of the two constituent learning domains, in particular the internet and the physical space. [2].

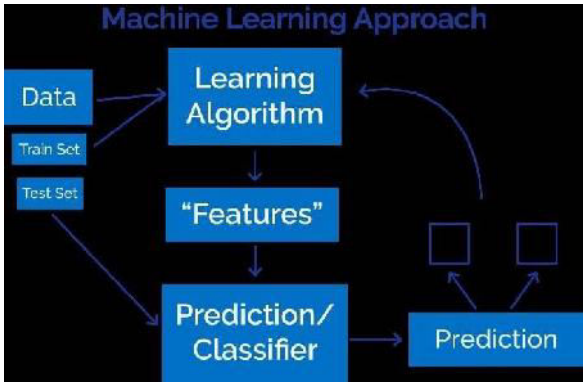
This argumentation appears on make right since majority of the data What's more communication science and technologies, on the person side, and physical framework science and technologies, on the other side, would epistemologically and methodologically different. The delegates of the multidisciplinary stance claim that the science technology about CPSs ought further synthesize the information and routines about the foundational physical, biological, building What's more data sciences, What's more if create a thorough science for CPSs. Those supporters of the trans-disciplinary elucidation case that once those science from claiming CPSs gives far extensive learning for implementation, the order if concentrate on giving requisition area free architectures What's more advances to fabricating useful cyber-physical artifacts and providing domain-orientated benefits [3].

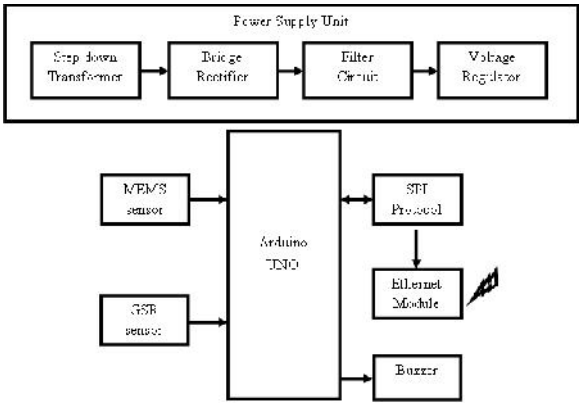
In our view, achieving all of these objectives can be considered as the mission of the science of CPSs. The objectives of the discipline of CPSs are:

- Mixing the information for different domains under a steady figure from claiming information with the goal as with underpin it perusing the fundamental standards about natural, formal, technical, social also human sciences.
- Creating a system-level understanding Furthermore theoretical frameworks from family of systems.

The principle Look into topics would for example, such that framework structure identification, advantageous interaction of physical and digital framework parts, combination from claiming empowering technologies, framework conduct analysis, self-sufficient system operation, ongoing framework control Also self-control, keen framework behavior, non-deterministic scenarios and protocols,

ISBN 978-93-88808-61-3





Also standards for next-generation usage. Likewise a whole, the order appears on a chance to be rather adolescent Also will be at present anguish starting with a to some degree unconsolidated, whether not confusing, wording. Test exploration On CPSs, and in addition prototyping-based testing would confronting experimental limits due to those vast scales, spatial distribution, inalienable complexity, prevailing heterogeneity and inserted nature. The idea and the term 'cyber-physical systems' popped dependent upon a percentage ten quite some time prior in the USA.

To Europe those same sort What's more manifestations for frameworks are named Possibly Likewise 'The Internet-of- Things'. [5], 'Web of Things', or as 'cooperative adaptive systems'.

Those expositive expression reflects An huge number for terms (such as, 'smart universal systems', 'deeply embedded systems', 'software-intensive systems', 'hybrid automata', sensor actuator networks, M2M (OECD), which attempt on indicate the same concept, setting accentuation once specific parts (Eg. Functionality, implementation, and applications) about complex frameworks that determinedly incorporate digital What's more physical parts. [6].

The utilization for different terms Eventually Tom's perusing Different scientists raises those inclination that they need aid working once totally distinctive field, However truth be told they deliver the same alternately fundamentally the same issues What's more aspects about CPSs. Hypothetical examination in this area about premium will beeven now really scattered What's more not streamlined. Actually, those expositive expression reveals to a huge number about models, also those mixture of reference frameworks. There would vast contrasts in the approaches, innovative work efforts, Also subsidizing projects in Europe, USA and Japan. The inspiration for our foundation investigate went starting with two perceptions. Our far reaching expositive expression study investigated that an expansive number about papers examines Also contributes to exactly particular parts about utilitarian frameworks, technologies, data flows, usage and requisitions of CPSs.

II. Methodology

A. Description

Machine learning algorithms differentiate into supervised or unsupervised. Supervised algorithms give both information also wanted output, furthermore with furnishing reaction something like those correctness about predictions throughout preparation. Over this, this procedure will apply which is nourished on new information. Unsupervised algorithms may be not necessity to prepare with fancied Conclusion information. Instead, they use an iterative methodology known as profound taking in will survey information also land at conclusions. Unsupervised learning algorithms only used for complex tasks than supervised learning systems.

The methods included over machine learning comparative to that of data mining and predictive modelling. Both obliges seeking through information will search for patterns and changing programme actions appropriately. Numerous individuals need aid great known In light of the use for machine learning in starting with shopping on the web Furthermore being served ads identified with their buy. This

ISBN 978-93-88808-61-3

REFERENCE PAPER	DESCRIPTION	ALGORITHMS USED	EVALUATION
Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems	• In this paper, An multi-layer interution detection model will be planned and created to attain effectiveness which improves the detection and classification rate accuracy. • Machine learning techniques (C5 decision tree, Multilayer Perceptron neural system Also Naïve Bayes) need been connected utilizing gain ratio to selecting the best Characteristics for each layer to utilize smaller storage space and get higher intrusion detection.	Machine Learning Approach like C5 decision tree, Multilayer Perceptron neural network and Naïve Bayes are used which produces High intrusion performance.	Only C5 eliminates False alarm rate for MLP. other algorithms face difficulty to eliminate False alarm rate for MLP
Machine learning-based CPS for clustering high throughputmachining cycle conditions	• In this paper, unsupervised machine learning algorithms in cyber-physical systems are the key features to work towards highly precise diagnosis tools. • In case of clustering techniques, the Gaussian mixture model is used and also provides optimal solution in terms of interpretation by machine tool experts. • The agglomerative hierarchical algorithm is used to determine cycle phases. K means as same as agglomerative hierarchical algorithm to inherit variables.	Gaussian mixture model. The agglomerative algorithm and K-means clustering are used.	No importance shown to upgrade CPS embedded electronics which enables the algorithm to implement on its FPGA.
Event Detection through Differential PatternMining in Cyber-Physical Systems	• In this paper, DP miner, an exhaustive data mining framework for using in wireless sensors which performs in a distributed and parallel manner and it is able to extract a pattern of sensors that have event information. • DPminer can greatly reduce the energy for computation and communication in the Cyber Physical Systems.	DP miner and Differential Sensor Mining technique is used.	Differential Sensor Mining Technique faces difficulty in analysing big data in MLP.

happens in light suggestion engines utilize machine learning in to identity test web promotion conveyance previously, very nearly constant. Likewise separated starting with the customize marketing, other as a relatable point machine learning in employments instances incorporate duplicity detection, spam filtering, system security risk detection, predictive upkeep Also fabricating news encourages.

Fig1. Block Diagram of Machine Learning Approach. For (supervised) classification and regression (the most common tasks):

- Algorithm selection: Choose an algorithm.
- Feature selection: Choose features that capture the important characteristics of the system.
- Training/model building: Use part of the labeled set to build the model
- Parameter optimization (cross validation): Optimize the parameters using a second part of the labeled set to minimize the error rate.
- Validation: Use the remainder of the dataset to validate and assess the performance of the tuned model
- Apply the Algorithm

EXAMPLE: "MEDICAL MONITORING- POST OPERATIVE WOUND ANALYTICS"

Patients after an operation usually go through the recovery/rehabilitation process where they follow a strict routine. That will do by using sensors.

After the major surgery as per instruction from surgerion patients should maintain a fixed position or else if the patients supposed to falls down. That the position level will be monitored (MEMS Sensor) .MEMS generally consists three position like x, y, z. If the changes will be in the position means that will be updated through web server. Because of this updating nurses or Ward in charge can get alert without direct monitoring.

GSR, standing for galvanic skin response, is a method of measuring the electrical conductance of the skin. Strong emotion can cause stimulus to your sympathetic nervous system. Due to this condition can able to know the Pain or stress level (GSR Sensor) which rose after involved in surgery will be viewed through the web page. Not only web page updation can give alert through buzzer also. If the sensor data is not received to the cloud means

the doctor or representative person cannot able to monitoring the patient health frequently. So that patient can be affected by unwanted pain or any other factors. So that our machine-learning approach will guide to rectify/ notify the problem like sensor failure, controller board failure, internet connection lost.

Conclusion

Thus in this survey we analyze several algorithms based on Machine Learning Approach in order to extract knowledge from sensor data at various platforms which performs critical piece in the data mining to figure out those occasion identification to Cyber-Physical frameworks (CPS) contrasting with differential sensor pattern (DSP) .we use machine learning algorithm for event detection where we implement C5, Decision Tree mining techniques etc where accurate predictive results are achieved and Anomaly detection, Gaussian Mixture model, agglomerative hierarchical algorithm and K-means clustering is surveyed for supporting big data analysis.

References

1. "Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems", Heba Ezzat Ibrahim etal International Journal of Computer Applications (0975 – 8887) Volume 56– No.7, October 2012
2. "Machine learning-based CPS for clustering high throughput machining cycle conditions", Javier Diaz-Rozoa,b*, Concha Bielzab, Pedro Larrañagab open access article 2017
3. "Event Detection through Differential PatternMining in Cyber-Physical Systems", Md Zakirul Alam Bhuiyan, Member, IEEE, Jie Wu, IEEE transactions on big data,2017
4. B. Sharma, I. Franjo, N.-M. Alexandru, C. Haifeng, and. Guofei, "Modeling and analytics for cyber-physical systems in the age of big data," SIGMETRICS Perform.
5. Evaluation Review,vol. 41, no. 4, pp. 74–77, 2014.

6. G. Spezzano and A. Vinci, "Pattern detection in cyber physical systems," Procedia Computer Science, vol. 52, no. 2015, pp. 1016–1021, 2015.O. Niggemann1, G. Biswas, J. S. Kinnebrew, H. Khorasgani,S. Volgmann, and A. Bunte,
7. "Data-driven monitoring of cyber physical systems leveraging on big data and the internet-of things for diagnosis and control," in Proceedings of the 26th 1International Workshop on Principles of Diagnosis, 2015, pp. 185–192.
8. J. Lee, H. D. Ardakani, S. Yang, and B. Bagheri, "Industrial big data analytics and cyber-physical systems for future maintenance and service innovation," Procedia CIRP, vol. 38, pp. 3–7, 2015.
9. J. Jara, D. Genoud, and Y. Bocchi, "Big data for cyber physical systems: An analysis of challenges, solutions and opportunities," in The 8th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2014, pp. 376–380.
10. Jensen, T. D., Jensen, F.V., and Nielsen. (2001). "Bayesian networks and decision graphs," Springer, Berlin.
11. Amor, N. B., Benferhat, S., and Elouedi, Z. (2003). Naive Bayesian networks in intrusion detection systems. In Proc. Workshop on Probabilistic Graphical Models for Classification, 14th European Conference on Machine Learning (ECML) and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD),Croatia.
12. Hall, M. A. (1999). "Correlation based Feature Selection for Machine Learning,". Available at: https://www.lri.fr/pierres/ donn% E9es/save/these/articles/lprqueue/ hall99correlationbased.pdf
13. Draper-Gil, G., Lashkari,A. H., Mamun, M.S. I., and Ghorbani, A. A.(2017). "Characterization of Encrypted and VPN Traffic using Tim erelated Features," in Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP 2016), 407–414.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

A Review on Data Preprocessing For Efficient Prediction in Customer Relationship Management

[1]Ganga Patur, Assistant Professor.
[2]D. Nilima Priyadarshini, Assistant Professor.
Malla Reddy College of Engineering,

ABSTRCT

CRM (ConsumerRelationship Management) is a customer-focused business strategy designed to optimize revenue, profitability, and customer loyalty. CRM can use information from outside or within a company allowing much better comprehension of its customers on the set basis or to your own foundation, by producing client personalized documents. An improved knowledge of the buyer's customs, pursuits and demands might grow the transaction. So, steady information regarding your clients' choices and preferences forms the cornerstone of productive CRM. Since organizations become internet (in other words, grow in to e business), the find it difficult to maintain faithfulness in their older customers and also to entice clients remains more crucial, as a competitor's enterprise internet site might be only 1 click away. In this paper we studied data prepossessing methods for client log data.

Keywords: Data prepossessing, log, competitor prediction and Big data.

INTRODUCTION

Voluminous of information active in those on-line World Wide Web have managed to get rather vital that you utilize automatic data mining and knowledge discovery procedures to learn person navigation tastes. The various manners of internet website usage using way of a specific user could possibly be detected with World Wide Web usage mining methods that can mechanically recover ordinary accessibility patterns employing the utilization of sooner user simply click flows utilized in weblog data files. These Programs might be properly used towards designing the internet page for your own user and also to encourage digital advertising. Net usage mining technologies incorporates methods from two hot search areas, specifically, data mining and also the World Wide Web. By assessing the competition understanding concealed in blogs,

internet usage mining may assist searchers to supply much better layout and enterprise worries to present much better navigation behavior. Many businesses are emphasizing buyer orientation to both maintain regular users to its growth of consumer relationship administration. Investigation of curious browsers, gives invaluable advice for internet site designer to swiftly react for their own unique wants. This chapter introduces the search methodology utilized to look exactly the upcoming page forecast approach.

CUSTOMER LOGDATA

Purchaser log info can be really a document that has tremendous sum of facts and by that data origin; lots of info abstractions might be generated. For example, page opinions, host periods, along with click-streams. In these abstracts, shared provisions and key words can be utilized as specified in Table 1. This portion in addition supplies an in-depth outline of this web-log document structure used from today's research work.

A log file will be understood to be a document which enrolls the surgeries of the internet server. Log data files returns advice such as for instance the data files which can be asked, sometime of this document ask the individual and also the speaking webpage. Every point of this log document defines one “strike" over the log file from your host plus comprises numerous subjects and also the arrangement of this log utilized for assesses change from host to host. Investigation of log document is Deemed valuable for the next reasons:

• The Internet server produces log documents, therefore getting an raw info is Not Too hard and Doesn't require any alterations or added programming attempt,

• Business's servers may maintain info inside their standard. It makes it possible to get a Institution to Alter applications after, utilize a lot Diverse applications and analyze chronological arrangement having a new program,

• Production and incorporating details for the log record doesn't need any extra Domain Name Server Look-ups. Ergo, There Aren't Any external server requirements which may slow down page loading rates, also Contributes to uncounted webpage viewpoints, and also

• The Internet Website's host documents all Trade it gets and this is considered reliable.

The arrangement of this log record is displayed at Table 1 & 2An hyphen ('-')

at one or more of these disciplines suggest missing info.

TABLE 1 IMPORTANT TERMS IN CUSTOMER LOG DATA

S.No.	Terms	Description
1	User	**Users accessing file from the web servers through a browser.**
2	Page View	**A page view is an abstract that consist of every file that is displayed on user's browser screen at one point of time. A page view may be associated with a single user action or can be related with several files such as scripts ,frames,and graphics, etc.,**
3	Hit	**Every successful file that is sent to the web browser is a hit**
4	Click Stream	**It is a sequential series of page view requests.**
5	Server Session or visit	**A Server Session or visit happens when a user or robot visits a website.**
6	User Session	**A user session is defined as a set of page requests made by a single user.**
7	**Customer Log**	**These are files that stores into them details regarding all the visits made to a web site or a portal automatically and are maintained in the web server.**

TABLE 2 CUSTOMER LOG FIL

S.No.	Name of Field	Description	Example value
1	IP Address	IP address of the Client who request for a page on the web server	**127.0.0.1**
2	UserID and Password	Provides the username and their corresponding password used during the access of a content-secured transaction	**Voder23 12ert35**
3	Timestamp	The date, time and time zone when the server finished processing the request.	**[10/Oct/2000:13:55:36 -0700]**
4	Access Request	Request line from the client. It has three parts, the METHOD, URL STEM and PROTOCOL used during transmission.	**GET http/www.yahoo.com/asctab31.zipHTTP/1.0**
	Method	Can be GET (request made to get a program or document) or POST (during transmission indicates the server that data is following) or HEAD (used by link checking programs, not browsers and downloads just the information in the HEAD tag information)	**GET POST HEAD**
	URL	The address of	**/download/windows/asctab31.zip**
	Protocol	**protocol**	**HTTP/1.0**

I. RESEARCH METHODOLOGY

Internet can be actually a client/server style and design by which a consumer sends an internet requests for within the web (WWW) into some internet server. The internet server reacts by reacting to this petition. The trade session includes the market of protocols and methods. But as a result of exponential increase of WWW, there really are a high quantity of customers that disagrees with all the servers with a high number of programs correlated with just one another, causing a significant raise the WWW latency and burden about the internet. If a proxy host set in between a web browser and a host, it's a effective tool which could possibly be utilized to decrease your WWW's latency. It follows that it may intercept any orders into the server to guarantee whether the request can be fulfilled by the client itself. If not, then it may be offered to the internet server. The clear presence of proxy servers also provides 2 major positive aspects as supplied just below.

- Reduce latency: Gradually, most of the asked consequences from several customers are saved inside a proxy-server.For example, contemplate if just two users and B get the web by means of a proxy host. Assume consumer A asks to get a specific webpage (P 1). Shortly after, consumer also requests for equal webpage. Without forwarding the petition for the internet server, then these pages is returned from your proxy host its own cache at which in fact the newly downloaded website pages have been kept. Considering proxy host and the consumer share exactly the Exact Same system, the surgeries are substantially quicker, thereby decreasing the perceived latency, and also
- Filter un-wanted Requests: Negative asks are all taken off from the Proxy servers. By way of instance, a faculty can confine the college students from obtaining a particular pair of the web sites using a proxy-server.

To reduce the WWW latency, the behavior of the consumer can be called and therefore the pages that are predicted are all pre-fetched and kept temporarily at the cache from their proxy host. The petition of this user could be fulfilled immediately when the webpage can be found from your cache. An overall site forecast version is displayed in Figure.1.

Web Requests

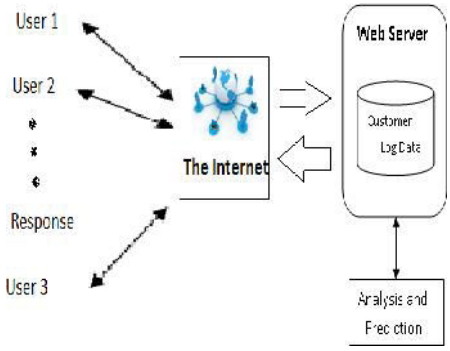


Figure1: General Architecture of Web System with Web Access Prediction

Due to person's successive actions within their communicating with the entire Internet presents a huge obstacle for investigators from the internet engineering field and also can be the primary focus with the exploration, forecast of person's foreseeable future asks is composed of varied endeavors and determine fig 1 offer the stream of those activities at the research job. The suggested strategy is known as adjoining page forecast approach. This job includes three major actions, particularly, pre-processing, competition consumer identification along with forecast of all future asks.

Inside this exploration function, every one of the aforementioned ways is taken care of like an individual period, which must be implemented at a sequential way throughout the plan and execution of internet site forecast procedure. The investigation methodology has been intended in a fashion that all measure tries to increase its individual endeavor and operates with all the intention of bettering its performance prediction. Throughout the stream of forecast, the outcome of one particular phase can be utilized as input signal the subsequent period. The suggested research frame is offered in Figure 3.3 along with the many processes enhanced throughout the plan of the next page forecast approach are all introduced at these sub sections.

Phase I: Prepossessing Algorithms

Prepossessing of a web log file is nothing but simply reformatting the entries of a log file into a form that can be used directly by the subsequent steps of the log analyzer.

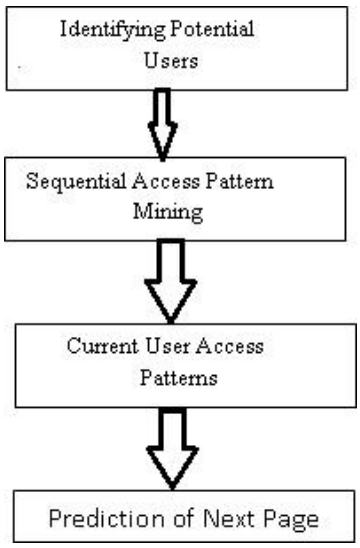


Figure 3.2: Tasks in next page prediction system

II. PREPROCESSING ALGORITHMS

The very first thing of this suggested second web-page forecast process will be per-processing, at which in fact the most important focus would be always to maintain simply applicable data out of the uncooked link. As a result of great number of insignificant data while in the internet log, the log may not be specifically utilized from the internet log mining treatment, thus at the prepossessing period, uncooked Internet logs will need to get cleaned, examined and changed to additional usage.

Period I of this analysis plays per-processing in 5 actions. They're recorded below and also the processes utilized in every measure are explained inside this chapter.

- Cleaning,
- User identification,
- Session identification,
- Formatting, and
- Clustering

III. CLEANING CUSTOMER LOG DATA

In the first step, that is, the task of cleaning raw web log data is considered. The data removed during cleaning are not required for user navigation and hence can be deleted safely from the log file. This step carries out the following tasks:-

- Removal of unwanted and redundant data,
- Removal of non-human accesses, and
- Removal of erroneous references.

Cases of undesirable data comprise asks including graphics, java script sand flash cartoons and video clip, etc. In case the file name contains gif, jpg, JPEG, CSS and so forth they are pruned out from the internet log document. Redundant statistics are recordings using similar

values in every single characteristic of this report. Instance of these statistics comprises admissions created by webmasters along with Spider accesses (instruments that scanning internet site to automatically extract its content). Search Engine normally utilizes system bots to creep throughout the web pages to get advice. The amount of information generated with these robots at a log record is high and has got a very poor impact whilst detecting navigation layout. This issue is solved inside thispaper by pinpointing the exact robot entrances first prior to devoting an individual collection in to rival and not-competitor end users. As stated by entrances from web-log produced with system robots can be identified by their IP address and agents. But this might require comprehension on most of form of representatives and see's, and this isn't easy to have. Another method will be to review the robots.txt document (positioned in the site's root directory), since a system convention has to read this document before obtaining the site. This really is due to the fact that the robots.txt gets got the access information of the site and every single robot will petition to learn its accessibility before scrawling. But that can't be relied on since obedience with robot exclusion standard is voluntary & the majority of the bots usually do not comply with exactly the suggested benchmark. So, to manually delete custom entrances, the next treatment issued.

Detect and remove all entries which has accessed robots.txtfile

Detect and remove all entries with visiting time of access as midnight (commonly used as the network activity at that time is light)

Remove entry when access mode is HEAD instead of GET orPOST

Compute browsing speed and remove all entries whose speed less than two seconds. The browsing speed is calculated as the number of viewed pages / session time.

CONCLUSION

In this paper we studied competitor prediction, in order to this first data per-processing is required, Real world data are generally Incomplete, Noisy and Inconsistent. Data cleaning, also called data cleansing or scrubbing. Fill in missing values, smooth noisy data, identify or remove the outliers, and resolve inconsistencies.Data cleaning is required because source systems contain "dirty data" that must be cleaned.In a customer relationship management (CRM) context, data prepossessing is a component of Web mining. Web usage logs may be pre-processed to extract

meaningful sets of data called user transactions, which consist of groups of URL references. User sessions may be tracked to identify the user, the Web sites requested and their order, and the length of time spent on each one. Once these have been pulled out of the raw data, they yield more useful information that can be put to the user's purposes, such as consumer research, marketing, or prediction.

REFERENCES

1. Ashish Bindra ;SrinivasuluPokuri ; Krishna Uppala ; Ankur Teredesai, 2012, "Distributed Big Advertiser Data Mining", ISSN: 2375-9232, 2012 IEEE 12th International Conference on Data Mining Workshops, PP: 914-914.
2. Abdel-Karim Al-Tamimi ; Raj Jain ; Chakchai So-In, 2010, "Dynamic resource allocation based on online traffic prediction for video streams", 2010 IEEE 4th International Conference on Internet Multimedia Services Architecture and Application, PP: 1---6.
3. BenleSu ;Yumei Wang ; Yu Liu, 2016, "Analysis and prediction of content popularity for online video service: a Youku case study", ISSN: 1673-5447, Volume: 13 , Issue: 12 , PP: 216-233.
4. ChengangZhu ;Guang Cheng ; Kun Wang, 2017, "Big Data Analytics for Program Popularity Prediction in Broadcast TV Industries", ISSN: 2169-3536, Volume: 5, PP: 24593-24601.
5. David K. Becker, 2017, "Predicting outcomes for big data projects: Big Data Project Dynamics (BDPD): Research in progress", 2017 IEEE International Conference on Big Data (Big Data), PP: 2320-2330.
6. Jian Ming ;Lingling Zhang ; Jinhai Sun ; Yi Zhang, 2018, "Analysis models of technical and economic data of mining enterprises based on big data analysis", 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), PP: 224-227.
7. Katarina Grolinger ; Miriam A.M. Capretz ; Luke Seewald, 2016, "Energy Consumption Prediction with Big Data: Balancing Prediction Accuracy and Computational Resources", 2016 IEEE International Congress on Big Data (BigData Congress), PP: 157-164.

8. KunZhang ;Minrui Fei ; Jianguo Wu ; Peijian Zhang, 2013, "Fast prediction model based big data system identification", 2013 Chinese Automation Congress, PP: 465-469.
9. Pedro Bastos ; Rui Lopes ; Luís Pires ; Tiago Pedrosa, 2009, "Maintenance behaviour-based prediction system using data mining", ISSN: 2157-3611, 2009 IEEE International Conference on Industrial Engineering and Engineering Management, PP: 2487-2491.
10. XiaojingMa ;Zhitang Li ; Hao Tu ; Bochao Zhang, 2010, "A Data Hiding Algorithm for H.264/AVC Video Streams Without Intra- Frame Distortion Drift", ISSN: 1051-8215, Volume: 20 , Issue: 10 , PP: 1320-1330.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

A Routing Protocol for Enhanced Efficiency in TEEN

Kota Divya Bharathi[1], Thirumala Vasala[2]
Assistant Professors, CSE Department,
Malla Reddy College of Engineering, Dhulapally, Kompally, Secunderabad-500100

Abstract

Wireless sensor networks are expected to find wide applicability and increasing deployment in the near future. In this paper, we propose a formal classification of sensor net- works, based on their mode of functioning, as proactive and reactive networks. Reactive networks, as opposed to passive data collecting proactive networks, respond immediately to changes in the relevant parameters of interest. We also in troduce a new energy efficient protocol, TEEN (Threshold sensitive Energy Efficient sensor Network protocol) for re- active networks. We evaluate the performance of our proto- col for a simple temperature sensing application. In terms of energy efficiency, our protocol has been observed to out- perform existing conventional sensornetwork protocols.

1. Introduction

In recent years, the use of wired sensor networks is being advocated for a number of applications. Some examples include distribution of thousands of sensors and wires over strategic locations in a structure such as an airplane, so that conditions can be constantly monitored both from the inside and the outside and a real-time warning can be issued when the monitored structure is about to fail.

Sensor networks are usually unattended and need to be fault-tolerant so that the need for maintenance is minimized. This is especially desirable in those applications where the sensors may be embedded in the structure or are in inhospitable terrain and are inaccessible for any service. The advancement in technology has made it possible to have extremely small, low powered devices equipped with programmable computing, multiple parameter sensing and wireless communication capability. Also, the low cost of sensors makes it possible to have a network of hundreds or thousands of these wireless sensors, thereby enhancing the reliability and accuracy of data and the area coverage as well. Also, it is necessary that the sensors be easy to deploy (i.e., require no installation cost etc). Protocols for these networks must be designed in such a way that the limited power in the sensor nodes is efficiently used. In addition, environments in which these nodes operate and respond are very dynamic, with fast changing physical parameters. The following are some of the parameters which might change dynamically depending on the application:

Power availability.

Position (if the nodes are mobile).

Reachability.

Type of task (i.e. attributes the nodes need to operate on)

So, the routing protocol should be fault-tolerant in such a dynamic environment. The traditional routing protocols defined for wireless ad hoc networks [1] [9] are not well suited due to the following reasons:

1. Sensor networks are "data centric" i.e., unlike traditionalnetworkswheredataisrequestedfromaspecific node, datais requestedbasedoncertainattributessuch as, which area has temperature $> 50\ F$?
2. The requirements of the network change with the application and so, it is application-specific [3]. For example, in some applications the sensor nodes are fixed and not mobile, while others need data based only on one attribute (i.e., attribute is fixedin this network).
3. Adjacent nodes may have similar data. So, rather than sending data separately from each node to therequesting node, it is desirable to aggregate similar data and send it.
4. In traditional wired and wireless networks, each node is given a unique id, used for routing. This cannot be effectively used in sensor networks. This is because, these networks being data centric, routing to and from specific nodes is not required. Also, the large number of nodes in the network implies large ids [2], which might be substantially larger than the actual data being transmitted.

This work is supported by the Ohio Board of Regents' Doctoral En- hancement Funds

Thus, sensor networks need protocols which are application specific, data centric, capable of aggregating data and optimizing energy consumption. An ideal sensor network should have the following additional features:

Attribute based addressing is typically employed in sensor networks. The attribute based addresses are composed of a series of attribute-value pairs which specify certain physical parameters to be sensed. For example, an attribute address may be (temperature *> 100 F* , location = ??). So, all nodes which sense a temperature greater than 100 *F* should respond with their location.

Location awareness is another important issue. Since most data collection is based on location, it is desirable that the nodes know their position whenever needed.

2. Related Work

In this section, we provide a brief overview of some related research work.

Intanagonwiwat et. al [7] have introduced a data dissemination paradigm called *directed diffusion* for sensor networks. It is a data-centric paradigm and its application to query dissemination and processing has been demonstrated in this work.

Estrin et. al [3] discuss a hierarchical clustering method with emphasis on localized behavior and the need for asymmetric communication and energy conservation in sensor networks.

A cluster based routing protocol (CBRP) has been proposed by Jiang et. al in [8] for mobile ad-hoc networks. It divides the network nodes into a number of overlapping or disjoint two-hop-diameter clusters in a distributed manner. However, this protocol is not suitable for energy constrained sensor networks in this form.

Heinzelman et. al [5] introduce a hierarchical clustering algorithm for sensor networks, called *LEACH*. We discuss this in greater detail in section 6.1.

3. Motivation

In the current body of research done in the area of wireless sensor networks, we see that particular attention has not been given to the time criticality of the target applications. Most current protocols assume a sensor network collecting data periodically from its environment or responding to a particular query. We feel that there exists a need for networks geared towards responding immediately to changes in the sensed attributes. We also believe that sensor networks should provide the end user with the ability to control the trade-off between energy efficiency, accuracy and response times dynamically. So, in our research, we have focussed on developing a communication protocol which can fulfill these requirements.

4. Classification of Sensor Networks

Here, we present a simple classification of sensor networks on the basis of their mode of functioning and the type of target application.

Proactive Networks

The nodes in this network periodically switch on their sensors and transmitters, sense the environment and transmit the data of interest. Thus, they provide a snapshot of the relevant parameters at regular intervals. They are well suited for applications requiring periodic data monitoring.

Reactive Networks

In this scheme the nodes react immediately to sudden and drastic changes in the value of a sensed attribute. As such, they are well suited for time critical applications.

5. Sensor Network Model

We now consider a model which is well suited for these sensor networks. It is based on the model developed by Heinzelman et. al. in [5]. It consists of a base station*(BS)*, away from the nodes, through which the end user can access data from the sensor network. All the nodes in the network are homogeneous and begin with the same initial energy. The *BS* however has a constant power supply and so, has no energy constraints. It can transmit with high power to all the nodes. Thus, there is no need for routing from the *BS* to any specific node. However, the nodes cannot always reply to the *BS* directly due to their power constraints, resulting in asymmetric communication.

This model uses a hierarchical clustering scheme. Consider the partial network structure shown in Fig. 1. Each cluster has a cluster head which collects data from its cluster members, aggregates it and sends it to the *BS* or an upper level cluster head. For example, nodes 1.1.1, 1.1.2, 1.1.3, 1.1.4, 1.1.5 and 1.1 form a cluster with node 1.1 as the cluster head. Similarly there exist other cluster heads such as 1.2, 1 etc. These cluster-heads, in turn, form a cluster with node 1 as their cluster-head. So, node 1 becomes a second level cluster head too. This pattern is repeated to form a hierarchy of clusters with the uppermost level cluster nodes reporting directly to the *BS*. The *BS* forms the root of this hierarchy and supervises the entire network. The main features of such an architecture are:

All the nodes need to transmit only to their immediate cluster-head, thus saving energy.

Only the cluster head needs to perform additional computations on the data. So, energy is again conserved.

ISBN 978-93-88808-61-3

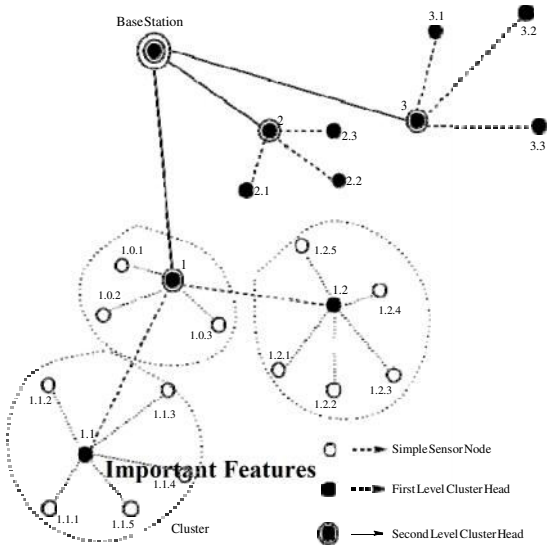


Figure 1. *HierarchicalClustering*

Cluster-heads at increasing levels in the hierarchy need to transmit data over correspondingly larger distances. Combined with the extra computations they perform, they end up consuming energy faster than the other nodes. In order to evenly distribute this consumption, all the nodes take turns becoming the cluster head for a time interval *T*, called the cluster period.

6. Sensor Network Protocols

The sensor network modeldescribed in section 5 is used extensively in the following discussion of sensor network protocols.

Proactive Network Protocol

In this section, we discuss the functionality and the characteristics expected in a protocol for proactivenetworks.

Functioning

At each cluster change time, once the cluster-heads are decided, the cluster-head broadcasts the following parameters :

***Report Time(T_R)*:** This is the time period between succes- sive reports sent by a node.

***Attributes(A)*:** This is a set of physical parameters which the user is interested in obtaining data about.

At every report time, the cluster members sense the parameters specified in the attributes and send the data to the cluster-head. The cluster-head aggregates this data and sends it to the base station or the higher level cluster-head, as the case may be. This ensures that the user has a com- plete picture of the entire area covered by the network.

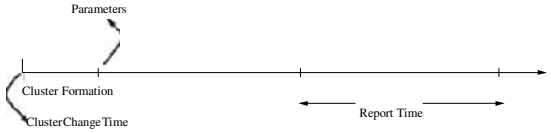


Figure 2. *Time line for proactive protocol*

The important features of this scheme are mentioned below:

1. Since the nodes switch off their sensors and transmitters at all times except the report times, the energy of the network is conserved.

2. At every cluster change time, T_R and A are transmitted afresh and so, can be changed. Thus, the user can decide what parameters to sense and how often to sense them by changing A and T_R respectively.

This scheme, however, has an important drawback. Because of the periodicity with which the data is sensed, it is possible that time critical data may reach the user only after the report time. Thus, this scheme may not be very suitable for time-critical data sensing applications.

LEACH

LEACH (Low-Energy Adaptive Clustering Hierarchy) is a family of protocols developed in [5]. LEACH is a good approximation of a proactive network protocol, with some minor differences.

Once the clusters are formed, the cluster heads broadcast a TDMA schedule giving the order in which the cluster members can transmit their data. The total time required to complete this schedule is called the frame time T_F . Every node in the cluster has its own slot in the frame, during which it transmits data to the cluster head. When the last node in the schedule has transmitted its data, the schedule repeats.

The *report time* discussed earlier is equivalent to the *frame time* in LEACH. The *frame time* is not broadcast by the cluster head, though it is derived from the TDMA schedule. However, it is not under user control. Also, the attributes are predetermined and are not changed midway.

Example Applications

This network can be used to monitor machinery for fault detection and diagnosis. It can also be used to collect data about temperature change patterns over a particular area.

Reactive Network Protocol: TEEN

In this section, we present a new network protocol called TEEN *(**T**hreshold sensitive **E**nergy **E**fficient sensor **N**etwork protocol)*. It is targeted at reactive networks and is the first protocol developed for reactive networks, to our knowledge.

Functioning

In this scheme, at every cluster change time, in addition to the attributes, the cluster-head broadcasts to its members,

Hard Threshold ***(H_T)*****:** This is a threshold value for the sensed attribute. It is the absolute value of the attribute beyond which, the node sensing this value must switch on its transmitter and report to its clusterhead.

Soft Threshold ***(S_T)*****:** This is a small change in the value of the sensed attribute which triggers the node to switch on its transmitter and transmit.

The nodes sense their environment continuously. The first time a parameter from the attribute set reaches its hard threshold value, the node switches on its transmitter and sends the sensed data. The sensed value is stored in an internal variable in the node, called the *sensed value (SV)*. The nodes will next transmit data in the current cluster period, only when *both* the following conditions aretrue:

1. The current value of the sensed attribute is greater than the hard threshold.

2. The current value of the sensed attribute differs from *SV* by an amount equal to or greater than the soft threshold.

Whenever a node transmits data, *SV* is set equal to the current value of the sensed attribute.

Thus, the hard threshold tries to reduce the number of transmissions by allowing the nodes to transmit only when the sensed attribute is in the range of interest. The soft threshold further reduces the number of transmissions by eliminating all the transmissions which might have otherwise occurred when there is little or no change in the sensed attribute once the hard threshold.

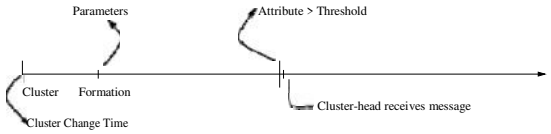


Figure 3. *Time Line for TEEN*

Important Features

The main features of this scheme are as follows:

1. Time critical data reaches the user almost instantaneously. So, this scheme is eminently suited for time-critical data sensing applications.

2. Message transmission consumes much more energy than data sensing. So, even though the nodes sense continuously, the energy consumption in this scheme can potentially be much less than in the proactive network, because data transmission is done less frequently.

3. The soft threshold can be varied, depending on the criticality of the sensed attribute and the target application.

4. A smaller value of the soft threshold gives a more accurate picture of the network, at the expense of increased energy consumption. Thus, the user can control the trade-off between energy efficiency and accuracy.

5. At every cluster change time, the attributes are broadcast afresh and so, the user can change them as required.

The main drawback of this scheme is that, if the thresholds are not reached, the nodes will never communicate, the user will not get any data from the network at all and will not come to know even if all the nodes die. Thus, this scheme is not well suited for applications where the user needs to get data on a regular basis. Another possible problem with this scheme is that a practical implementation would have to ensure that there are no collisions in the cluster. TDMA scheduling of the nodes can be used to avoid this problem. This will however introduce a delay in the reporting of the time-critical data. CDMA is anotherpossible solution to this problem.

Example Applications

This protocol is best suited for time critical applications such as intrusion detection, explosion detection etc.

7. Performance Evaluation

Simulation

To evaluate the performance of our protocol, we have implemented it on the ns-2 simulator [10] with the *LEACH* extension [4]. Our goals in conducting the simulation are as follows:

Compare the performance of the TEEN and LEACH protocols on the basis of energy dissipation and the longevity of the network.

Study the effect of the soft threshold ST on TEEN.

The simulation has been performed on a network of 100 nodes and a fixed base station. The nodes are placed randomly in the network. All the nodes start with an initial energy of 2J. Cluster formation is done as in the *leach* protocol [5] [6]. However, their radio model is modified to include idle time power dissipation (set equal to the radio electronics energy) and sensing power dissipation (set equal to 10% of the radio electronics energy). The idle time power is the same for all the networks and hence, does not affect the performance comparison of the protocols.

Simulated Environment

For our experiments, we simulated an environment with varying temperature in different regions. The sensor network nodes are first placed randomly in a bounding area of 100x100 units. The actual area covered by the network is then divided into four quadrants. Each quadrant is later assigned a random temperature between $0\ F$ and $200\ F$ every 5 seconds during the simulations. It is observed that most of the clusters have been well distributed over the four quadrants.

Experiments

We use two metrics to analyze and compare the performance of the protocols. They are:

Average energy dissipated: This metric shows the average dissipation of energy per node over time in the network as it performs various functions such as transmitting, receiving, sensing, aggregation of data etc.

Total number of nodes alive: This metric indicates the overall lifetime of the network. More importantly, it gives an idea of the area coverage of the network over time.

We now look at the various parameters used in the implementation of these protocols. A common parameter for both the protocols is the attribute to be sensed, which is the temperature.

The performance of TEEN is studied in two modes, one with only the hard threshold *(hard mode)* and the other with both the hard threshold and the soft threshold *(soft mode)*. The hard threshold is set at the average value of the lowest and the highest possible temperatures, $100\ F$. The soft threshold is set at $2\ F$ for our experiments.

Results

We executed 5 runs of the simulator for each protocol and for each mode of TEEN. The readings from these 5 trials were then averaged and plotted. A lower value of the energy-dissipation metric and a higher number of nodes alive at any given time indicates a more efficient protocol.

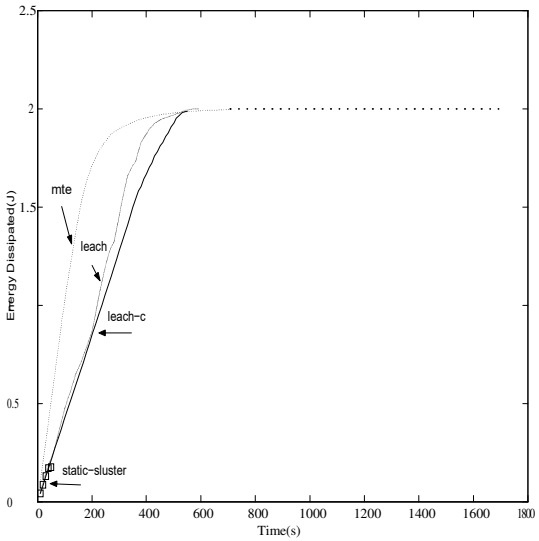


Figure 4. *Energy dissipation: LEACH*

Figures 4 and 5 show the behavior of the network in proactive mode. This comparison was originally done in LEACH [6]. It is repeated here taking into account the modified radio energy model. Of the four protocols [6], *mte (minimum transmission energy)* lasts for the longest time. However, we observe from Fig. 5 that only one or two nodes are really alive. As such, *leach* and *leach-c* (a variant of *leach*) can be considered the most efficient protocols, in terms of both energy dissipation and longevity.

In Figures 6 and 7, we compare the two protocols. We see that both modes of TEEN perform much better than *leach*. If the cluster formation is based on the *leach-c* protocol, the performance of the TEEN protocol is expected to be correspondingly better.

As expected, *soft mode* TEEN performs much better than

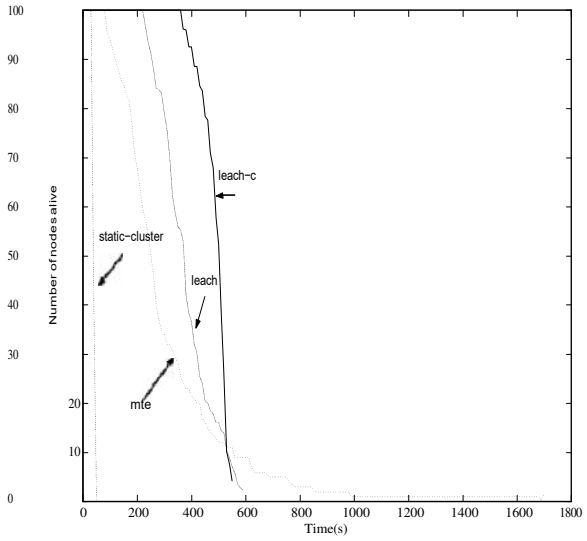


Figure 5. *No. of nodes alive: LEACH*

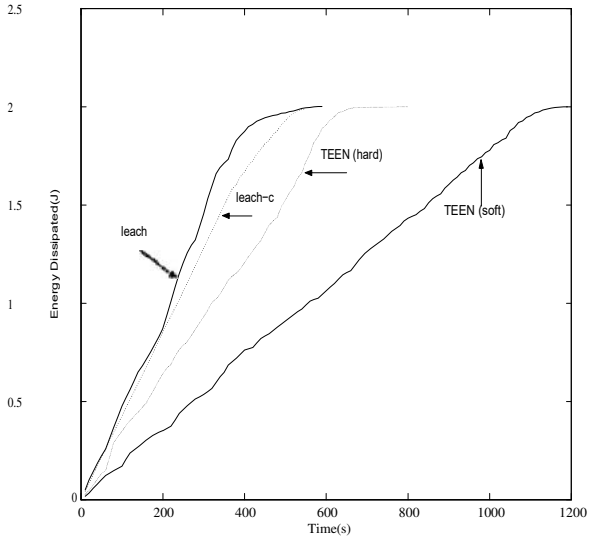


Figure 6. *Comparison of average energy dissipation*

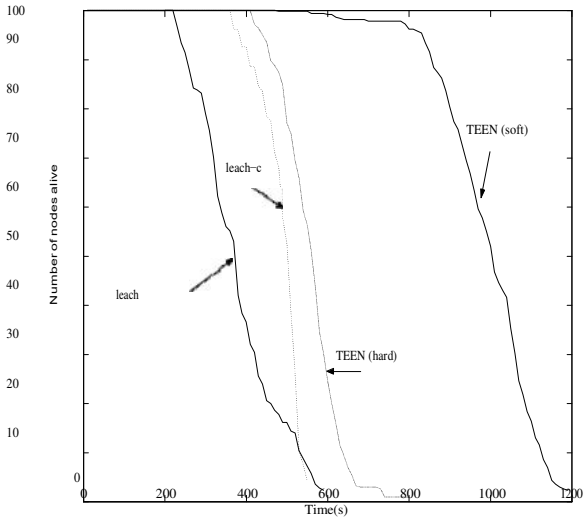


Figure 7. *Comparison of the no. of nodes alive*

hard mode TEEN because of the presence of the soft threshold.

8. Conclusions

In this paper, we present a formal classification of sensor networks. We also introduce a new network protocol, *TEEN* for reactive networks. TEEN is well suited for time critical applications and is also quite efficient in terms of energy consumption and response time. It also allows the user to control the energy consumption and accuracy to suit the application.

Acknowledgment

We would like to thank Wendi Heinzelman for her valuable suggestions and for letting us use her *LEACH* extension to the *ns* simulator for our experiments.

References

[1] J. Broch, D. Maltz, D. Johnson, Y. Hu, and J. Jetcheva. "A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols". In *Proceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking(MOBICOM), ACM*, October 1998.

[2] J. Elson and D. Estrin. "An Address-Free Architecture for Dynamic Sensor Networks". Technical Report 00-724, Computer Science Department, USC, January 2000.

[3] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar. "Next Century Challenges: Scalable Coordination in Wireless Net- works". In *Proceedings of the 5th Annual ACM/IEEE In- ternational Conference on Mobile Computing and Network- ing(MOBICOM)*, pages 263–270, 1999.

[4] W. Heinzelman, A. Chandrakasan, and H. Balakrish- nan. *"uAMPS ns Code Extensions"*. http://www-mtl.mit.edu/research/icsystems/uamps/ leach.

[5] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan. "Energy-Efficient Communication Protocols for Wireless Microsensor Networks". In *Proceedings of Hawaiian In- ternational Conference on Systems Science*, January 2000.

[6] W. B. Heinzelman. *"Application-Specific Protocol Archi- tectures for Wireless Networks"*. PhD thesis, Massachusetts Institute of Technology, June 2000.

[7] C. Intanagonwiwat, R. Govindan, and D. Estrin. "Di- rected Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks ". In *Proceedings of the 6th Annual ACM/IEEE International Conference on Mobile Computing and Networking(MOBICOM)*, pages 56–67, Au- gust 2000.

[8] M. Jiang, J. Li, and Y. C. Tay. "Cluster Based Routing Pro- tocol". Internet Draft, 1999.

[9] E. M. Royer and C.-K. Toh. "A Review of Current Rout- ing Protocols for Ad-Hoc Mobile Wireless Networks". In *IEEE Personal Communications Magazine*, pages 46–55, April 1999.

[10] UCB/LBNL/VINT. *"Network Simulator-ns"*. http://www- mash.cs.berkeley.edu/ns.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Efficient Irrigation System using IOT And Raspberry Pi

Puladas Sandhya Priyanka[1] , K.John Bunyan[2]

[1] *Assistant Professor, Dept. of CSE, Malla Reddy College of Engineering, Secunderabad*
[2] *Assistant Professor, Dept. of CSE, Malla Reddy College of Engineering, Secunderabad*

Abstract

Water is the important source in human life. Around 80 % to 90 % water used in agriculture field. As due to day by day growth in globalization and population water consumption is also increases. There is challenge in front of every country to reduce the farm water consumption and provide fresh and healthy food. Today automation is one of the important role in human life. The system is not only provides comfort but also reduce energy, efficiency and time saving. Whenever there is a change in temperature, humidity and current status of rain of the surroundings these sensors senses the change in temperature and humidity and gives an interrupt signal to the raspberry pi. Now a day the industries are using an automation and control machines which are high in cost and not suitable for using in a farm & garden field. So in this work we design a smart irrigation technology based on IOT using Raspberry pi. The system can be used to control the water motor automatically and can also monitor the growth of plant by using webcam. We can watch live streaming of farm on mobile phone using suitable application by using Wi-Fi network. Raspberry pi is the main heart of the overall system.

Key Words*: Raspberry Pi, Wi-Fi, Sensors, IOT, automation*

I. INTRODUCTION

India is one of the largest freshwater user in the world, and our country uses large amount of fresh water than other country. There is a large amount of water used in agriculture field rather than domestic and industrial sector. 65% of total water is contributes as a groundwater. Today water has become one of the important source on the earth and most of used in the agriculture field. As the soil-moisture sensor and temperature sensor are placed in the root zone of the plants, the system can distributed this information through the wireless network. The raspberry pi is the heart of the system and the webcam is interfaced with Raspberry pi via Wi-Fi Module. Python programming language is used for automation purpose. The system is a network of wireless sensors and a wireless base station which can be used to provide the sensors data to automate the irrigation system. The system can used the sensors such as soil moisture sensor and soil temperature sensor and also ultrasonic sensor. The raspberry pi model is programmed such that if the either soil moisture or temperature parameters cross a predefined threshold level, the irrigation system is automated, i.e. the relay connected to the raspberry pi will turn ON or OFF the motor. This paper present an efficient, fairly cheap and easy automated irrigation system. This system once installed it has less maintenance cost and is easy to use. By using the webcam with suitable application on mobile phone we can easily online monitoring the actual situation of the field and sensors such as soil moisture and temperature are used to provide the information about changes occurs in the field. It is more advantageous than the traditional agriculture techniques.

II. RELATED WORK

After extensive research in the agricultural field, many researchers found that the agriculture area and its productivity are decreasing by the day. With the Use of different technology in the field of agriculture we can increase the production as well as reduce manual efforts. This paper shows the technology used in agriculture sector based on IOT and Raspberry Pi. Chandan kumar Sahu proposed a system on "A Low Cost Smart Irrigation Control System". It includes a number of wireless sensors which are placed in different directions of the farm field. Each sensor is integrated with a wireless networking device and the data received by the "ATMEGA318" microcontroller which is on the "ARDUINO-UNO" development board. The Raspberry pi is used to send various types of data like text messages and images through internet communication to the microcontroller process [1]. Supraha Jadhv proposed, automated irrigation system using wireless sensor network and raspberry pi that control the activities of drip irrigation system efficiently [2]. Sebastian Hentzelt proposed a paper on the water distribution system and gave results to decompose the original nonlinear optimal control problem (OCP) [3]. Joauin Gutierrez attempted a paper that research automated irrigation system using a wireless sensor network and GPRS module instead of the Raspberry pi [4]. Ms. Deweshvree Rane Proposed "Review paper based on Automatic Irrigation System Based on RF Module" it is based on the RF module, this device is used to transmit or received radio signal between two devices. It's design is complex because of the sensitivity of radio circuits and the accuracy of the components [5]. Karan Kansara proposed "Sensor based automatic irrigation system with IoT", this irrigation system is used a rain gun pipe, one end connected to the water pump and another to the root of plant. It doesn't provide water as a natural rainfall like sprinkler and also it uses only soil moisture sensor [6]. G. Parameswaran proposed "Aurdino based smart irrigation system using Internet of Things", the researcher has not used Raspberry pi instead the work is done using aurdino controller without use of soil moisture sensors [7].

III. PROPOSED SYSTEM

The block diagram of the proposed system is as shown in the below figures. Fig a. represents the Transmitting Section whereas Fig b. represents the Receiving Section. The main components of this diagram are Sensors, Raspberry Pi module, Wi-Fi connection, LDR, relay, motor, and lamp.

ISBN 978-93-88808-61-3

Block Diagram

a. Transmitting section

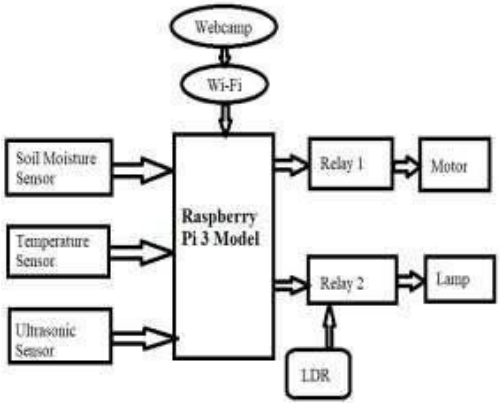


Fig a: Irrigation Control System (Transmitting Section)

The above figure shows that main block diagram of Irrigation control system. In that main model is Raspberry pi 3 model, Relays, LDR, Sensors. In this control system three sensors are such as soil moisture sensor, temperature sensor, ultrasonic sensors are connected to the raspberry pi 3 model also Wi-Fi connection is connected to the model. The connection of raspberry pi is given to the relay 1 and relay 2 which are again given to the motor and lamp respectively. LDR connection is given to the relay 2.

b. Receiving section

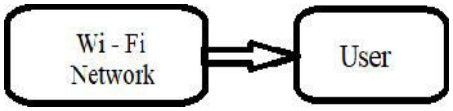


Fig b: Monitoring Unit (Receiving Section)

Above figure shows that receiving section of the main module i.e Monitoring unit. The two sections present are: one is Wi-Fi network and user. This connection again given to the raspberry pi 3 module.

SENSORS

A sensor is a device, module, or subsystem whose purpose is to detect events or changes in its environment and send the information to other electronics, frequently a computer processor. In short sensors are the device which converts the physical parameter into the electric signal. A sensor's sensitivity indicates how much the sensor's output changes when the input quantity being measured changes. The system which shown in fig.1 consists of

- Soil moisture sensor- used to measure the moisture content of the soil.
- Temperature sensor - used to detect the temperature of the soil.
- Ultrasonic sensor - used to measure the water level in the water tank.

Fig 3.2.1: Soil Moisture Sensor

Fig 3.2.2: DHT 11(Temperature Sensor)

RASPBERRY PI

Raspberry Pi is a small sized single board computer which is capable of doing the entire job that an average desktop computer does like spread sheets, Word processing, Internet, Programming, Games etc. It contain 1GB RAM, 2 USB, ARM V8 Processor and an Ethernet port, HDMI & RCA ports for display,3.5mm Audio jack, SD card slot (bootable), General purpose I/O pins, runs on 5v.

Fig 3.3: Raspberry Pi Model

RELAY

A relay is an electrically operated switch. Relays are used where it is necessary to control a circuit by a separate low-power signal. A relay with calibrated operating characteristics and sometimes multiple operating coils are used to protect electrical circuits from overload. As shown in above figure raspberry pi is connected to the devices via relay. Here relay can be operated as switch to on or off the devices.

Fig 3.4: Relay

IV. WORKING PRINCIPLE

As the Raspberry Pi is the heart of the system. This system contain webcam which is interfaced to Raspberry Pi via Wi- Fi module. The Raspberry Pi Model zero incorporates a number of enhancements and new features. This features of raspberry pi are improved power consumption, increased connectivity and greater IO which made this powerful, small and lightweight ARM based computer. The Raspberry Pi cannot directly drive the relay. It has only zero volts or 3.3 V. It needs 12V to drive electromechanical relay. In that case it uses a driver circuit which provides 12V amplitude to drive the relay. Various sensors are connected to the Raspberry Pi board give a resistance variation at the output. This output signal is applied to the comparator and signal conditioning circuit which has potentiometer to decide the moisture level above which the output of comparator goes high. This output signal is given to the Raspberry Pi board. If the soil moisture value is above the moisture level then the 3 phase induction motor will be OFF, whereas if the moisture level is low motor will be ON through the relay. LDR (Light Dependent Resistor) is used to control the light automatically and by using this we can monitor the farm at night also.

V. WORK FLOW OF THE SYSTEM

Step 1: Start.

Step 2: The system can be initialize on Raspberry Pi.

Step 3: The water level sensor constantly checks for the water level of the motor.

Step 4: The soil moisture sensor checks the soil moisture level constantly.

Step 5: The USB camera installed with the Raspberry Pi gives the complete lookout of the field and this can be monitored in the internal network system.

Step 6: The sensor constantly senses the temperature and humidity of the field and updates the date in the web server.

Step 7: If the permissible level of water is reduces, then the relay which is connected to the Raspberry Pi will turn ON the motor.

Step 8: Similarly, if the soil becomes dry, the motor which is connected to the relay will be turned ON to wet the field.

Step 9: If the step 8 is completed, it will go to the step 4.

Step 10: Similarly, if the step 7 is over, the command will go to the step 3.

VI. HARDWARE PART AND RESULT

ISBN 978-93-88808-61-3

Fig: Hardware part and result shown on android APP

VII. CONCLUSION

The smart irrigation system is suitable and cost effective for advance water resources for agricultural production. The system would provide feedback control system which will monitor and control all the activities of plant growth and irrigation system efficiently. If rain gun sensor can be added so that when it rains there won't be floods. Rain water harvesting can be done and this harvested water can be used to irrigate fields. We can also include many more water quality sensors that affect the crops.

VIII. REFERENCES

1. Chandankumar Sahu, Pramitee Behera. 2015 A Low Cost Smart Irrigation Control System IEEE Sponsored 2nd International Conference on Electronics and Communication System (ICECS).

2. Suprabha Jadhav, Shailesh Hambarde. 2013 Automated Irrigation System using Wireless Sensor Network and Raspberry pi, International Journal of science and research (IJSR).

3. Gutierrez, J. Francisco, J. Villa-Medina Nieto-Garibay, A., and Angel, P.G. 2013. Automated Irrigation System Using a Wireless Sensor Network and GPRS Module

4. Rane D, Indurkar P, and Khatri, D.M. 2015 Paper based on Automatic irrigation system on RF module In IJAICT Volume 1, Issue 9.

5. Kansara, K., Zaveri, V., Shah, S., Delwadkar, S., Jani, K. 2015 Sensor Based Automated Irrigation System with IOT In IJCSIT, Vol. 6.

6. Parameswaran, G., Sivaprasath, K. 2016 Arduino Based Smart Drip Irrigation System Using Internet Of Things In: IJESC, Volume 6 Issue No. 5.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Text Mining and Exploration using SVM

Arun.k[1] Syed Mohammed Shafi[2]
[1] Mallareddy college of Engineering, CSE dept, HYD
[2] Mallareddy college of Engineering, CSE dept,HYD,

Abstract— **On the basis of analyzing the basic concepts and the process of text excavation, the present study proposes some new methods in extraction of text features, deflation of characteristic collection, extraction of study and knowledge pattern, and appraisal of model quality. Meanwhile, it makes a comparison of two types of text categorization, text classifications and text cluster, and it briefly explores the basic issues to be solved in the future development of the text excavation technology.**

Key words- Text excavation; Text Features; characteristic collection deflation; Text classification; Text cluster

Ⅰ. INTRODUCTION

Along with the Internet application's popularization, Web already developed into has 300,000,000 page's distributional information spaces, moreover this digit still by the speed which every half year doubles grew. In the middle of these mass data, the majority of information are the non-structurization perhaps half structurization, moreover is containing the huge potential value knowledge. The people urgent need can from Web fast, discover these valuable knowledge effectively. On Web the information multiplicity has decided the Web knowledge discovery multiplicity. According to the processing object's difference, may the Web knowledge discover that divides into two broad headings: Content discovery and structure discovery. The content discovered that is mainly the excavation which keeps off to article this article. The text excavation (TextMining), may the massive documents set content carry on the abstract, classified, the cluster, the connection analysis as well as to Web on carries on the tendency forecast to the documents and so on.

Ⅱ.BASIC CONCEPTS

The text is by the massive characters, the word, the sentence is composed, to text excavation, in paramount consideration text character word. In English, Chinese and so on the natural language, have the massive words the concurrently kind of phenomenon, this for the text part-of-speech tagging, semantic labelling has brought the very major difficulty. Therefore,how to remove the part of speech, the semantic different meanings, is the text automatic labelling research key question.

A. A part-of-speech taggings

Ⅰ*)Concurrently kind of word:* Has two or two above lexical category glossary calls the concurrently kind of word.

the concurrently kind of word displays the different semantics in the different context linguistic environment, is by the concurrently kind of word lexical category decided that this is in the text excavation part-of-speech tagging question.

Concurrently kind of word classification Same-type opposite sex different righteousness concurrently kind of word

For example: Chairman Mao leads us to fight for state power. (“leadership” is a verb, leads, meaning of the instruction)

Chairman Mao is our good leadership. (“leadership” of is noun, meaning of person in charge, the leader)

Same-type opposite sex synonymy concurrently kind of word

For example: He has worked for 3 hours. (“hour” is classifier, Unit of time)

We measure the operating time by the hour. (“hour” is noun, Unit of time)

Heterogeneous homogeneous synonymy concurrently kind of word

For example: The computer has bought 50 computers. (“computer” is noun and “computer” synonymy)

The computer has bought 50 computers. (“computer” is noun and “computer” synonymy)

The non-word usage (stops word usage): In text relatively auxiliary functional word.

Ⅱ)*Non-word usage classification*

Function word: In English “a, the, for, with,…”; In Chinese “,…”And so on.

Full word: In database conference’s paper “database” a word, although the frequency of use is very high, but regards as the non-word usage.

Ⅲ *) Stem question:*compute, computes, computed identifies a word (distortion).

Ⅳ*) part-of-speech tagging:* The so-called part-of-speech tagging is for the text in word labelling part

of speech. Is mainly refers to the concurrently kind of word the lexical category to determine that the concurrently kind of word's lexical category determines only the sentence in according to the context.

B semantic labelling

semantics labelled a word to be equivocal, has formed the word different meanings phenomenon, semantic labelling mainly solves the word different meanings problem. A word equivocal is also in the natural language common phenomenon, but, in certain context, a word can only explain that generally is one semantics.

semantics labelling is to appears the words and expressions semantics carries on the determination in certain context, determined that its correct semantics and labels.

Semantic automatic labelling method Usual word is composed of meaning

The related word's method which appears using the retrieval context in determines the polysemant righteousness item

Determines the polysemant using the context matching relations the word meaning

To dispel equivocally with the most greatly possible righteousness item

C labelling technologies

The commonly used labelling technology route is based on the probability statistics and based on the rule method.

I *)Based on probability statistics CLAWS algorithm*

CLAWS is English Constituent-Likelihood Automatic Word-tagging System (ingredient likelihood automatic lexical category automatic labelling system) one algorithm which the abbreviation, it was in 1983 Ma Shaer (Mashall) when gives the LOB corpus (to have each literary style British English corpus, storage capacity quantity is 1,000,000 words) made automatic part-of-speech tagging proposed

II*) Based on probability statistics VOLSUNGA algorithm*

The VOLSUNGA algorithm is to the CLAWS algorithm improvement, in optimal path's choice aspect, is not only then calculates the probability to accumulate the biggest mark string finally, but along direction from left to right, the use “fortifies at every step” the strategy, regarding the current consideration's word, only retains leads to this word the optimal path, discards other ways, then embarks again from this word, carries on the match this way with next word's all marks, continues to discover the best way, discards other ways, goes forward like this gradually, walks until the entire cross section, obtains the entire cross section the optimal path to take the result output. Counts each word according to the corpus the relative labelling probability (Relative Tag Probability), and is auxiliary the optimal path with this kind of relative labelling probability the choice. The VOLSUNGA algorithm reduced the CLAWS algorithm time order of complexity and the spatial order of complexity greatly, raised the automatic part-of-speech tagging rate of accuracy.

The CLAWS algorithm and the VOLSUNGA algorithm are based on the statistical automatic labelling method, acts according to merely with the present probability labels the lexical category. But, with the present probability is only the biggest possibility and is not the only possibility, by determines the concurrently kind of word with the present probability, is by discards with the present probability low possible premise. In order to enhance the automatic part-of-speech tagging the accuracy, but must auxiliary by based on the rule method, determines the concurrently kind of word according to the language rule.

D other text retrieval labelling technology

I *) Inverted index*

Inverted index is an index structure that contains two hash tables index table, or two B +-tree index table, shown in Table 1, Table 2.

Table 1 Document Table (document_table)

doc_ID	posting_list
Doc_1	t_1_1, ..., t_1_n
Doc_2	t_2_1, ..., t_2_n
⋮	⋮
Doc_n	t_n_1, ..., t_n_n

Table 2 vocabulary (term_table)

term_ID	posting_list
Term_1	doc_1, ..., doc_i
Term_2	doc_1, ..., doc_j
⋮	⋮
Term_n	doc_1, ..., doc_n

Table 1 is composed of a group of documents record, posting_list is appears in the documents the word tabulation; Table 2 are composed of a group of word record, posting_list is contains this word the documents marking tabulation. Through such two tables, may discover with and assigns the documents related all words as well as with group of word related all documents for the decisive remark collection related all documents. But cannot process the synonym and the polysemant question, and posting_list is long, causes the memory expenses to increase.

II*) Signature File*

Features file is a storage database, the characteristics of each record of a document file. A feature of each bit corresponds to a fixed-length string, a bit corresponds to a word, if a particular word corresponds to appear in the document is, then the location of one, otherwise set 0.

IIITEXT EXCAVATION PROCESS

The text excavation object usually is group of HTML perhaps the XML form documents collection. Text excavation's general treating processes is: Document Set,Characteristics of the establishment of,Reduced feature set, Learning and knowledge extraction model,

Model Quality Evaluation-Knowledge model.

ISBN 978-93-88808-61-3

A . Text Features

Text feature refers to the metadata on the text. It can be divided into descriptive features (text, name, date, size, type, etc.) and semantic features (text, author, title, organization, content, etc.). Text feature to feature vectors, said: , Where t_i for the entry entry, $w_i(d)$ for t_i in d in the weights. As the feature vector dimension is usually very high, generally use the evaluation function to carryout feature selection. Evaluation of commonly used functions: nformation Gain,Expected Cross Entropy, Mutual Information,the Weight of Evidence for Text, Word Frequency.

Document Modeling：Using vector space model (VSM) of the text document model.

Frequency Matrix: line corresponds to the word *w,* the column vector corresponding to the document *d,* the simplest vector of values of words in the document appears on the value of 1, otherwise value is 0, Table 3 is based on occurrences of the word for the word frequency vector matrix, the value to reflect the word *w* and a document *d* of the correlation.

Table 3 Frequency of the Frequency Matrix document

	d_1	d_2	d_3	d_4	d_5	d_6
w_1	322	85	35	69	15	320
w_2	361	90	76	57	13	370
w_3	25	33	160	48	221	26
w_4	30	140	70	201	16	35

With the similarity of the document word frequency matrix can be measured, the typical method is the cosine similarity metric calculation (Cosine Measure).

Cosine similarity definition: $SIM(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|}$, is two documents vectors, The inner product is the standard vector dot product, Defines for $\sum_{i=1}^{w} v_{1i} v_{2i}$, is defined for $\sqrt{v_1 \cdot v_1}$

B characteristic collection deflation

Term frequency matrix similarly Gao Weishu, sparse data influence, to overcome these questions, the people proposed the latent semantic index (Latent Semantic Indexing) the method reduces the characteristic collection.

I) *Latent semantic index* ： “The singular value decomposes (Singular Value Decomposition using the matrix theory, SVD)” the technology, transforms the term frequency matrix as the singular matrix (K×K), concrete step:

- Establishment term frequency matrix, frequency matrix.

  Calculates frequency matrix the singular value decomposition. Decomposes frequency matrix to become 3 matrix U, S, V. U and V is the orthogonal matrix (UTU=I), S is the singular value diagonal matrix (K×K).

  Regarding each documents d, after removing in SVD eliminates the word new vector replace original vector.

  Preserves all vector set, founds the index with the high-level multi-dimensional index technology for it.

  Carries on the similarity computation after the transformation documents vector.

C studies and knowledge pattern extractions

I) *Participle*： The participle refers to between the text word and the word adds on the blank space, refers to Chinese text, because between English itself word and the word is differentiates by the blank space.

II) *Automatic participle*： The automatic participle is refers to uses the computer adds on the blank space automatically between the word and the word. The use is:

a) Chinese text automatic retrieval, filtration.
b) Classification and abstract.
c) Chinese text automatic proofreading.
d) Outside Chinese machine translation.
e) Chinese character recognition.
f) Chinese speech synthesis.
g) Take sentence as unit's Chinese character keyboard entry.
h) Chinese character Jan traditional form transformation.

III) *Main participle method*

a) Biggest match law (Maximum Matching method, MM law): The selection contains 6-8 Chinese characters the strings to take the biggest string, matches in the biggest string and the dictionary word clause, if cannot match, slices off a Chinese character to continue to match, until found the corresponding word in the dictionary. The match direction is from right to left.

b) Reversion biggest match law (Reverse Maximum method, RMM law): The match direction and the MM law are opposite, is from left to right. The experiment indicated: Regarding Chinese, the reversion biggest match law is more effective than the biggest match law.

c) Bilateral matching law (Bi-direction Matching method, BM law): Compared with the MM law and RMM law participle result, thus decides the correct participle.

d) Optimum matching law (Optimum Matching method, OM law): The dictionary in word according to them in the text appearance frequency's size arrangement, the high frequency's word arranges before, the frequency low word arranges, thus enhancement match speed.

e)Association backtracking: Uses the mechanism which associates and recalls to carry on the match.

IV)*Feature extraction*

ISBN 978-93-88808-61-3

The feature extraction is the glossary which, the phrase feature extraction appears to the text.

Characteristic weight function:

$$f_w(t_i) = \frac{f_v(t_i)\log(1+f_v(t_i))}{\sqrt{\sum_{j=1}^{m}(f_v(t_j)\log(1+f_v(t_j)))^2}}$$

And: Expresses the characteristic weight function; Expresses the characteristic item in the text frequency; Expressed that the characteristic paragraph frequency, namely contains ti the paragraph number/text total paragraph number.

V)*Automatic digest*

The automatic digest is uses the computer to withdraw automatically from the primitive documents reflects this documents center content accurately comprehensively the simple coherent short written work. Our country in 1995 carried on to the automatic digest system has evaluated, the system which for the first time participated has 3. The evaluation result performance is:

1. Three systems may according to the ratio which assigns from the original text select part of sentences.
2. The extraction digest is in the original text sentence, only then in the system 2 digests has rejected some digit.
3. Three system's digests do not superpose nearly completely.

may see the automatic digest system from above result also to have many foundation work to do. the text abstract is refers to from the documents extracts the key information, carries on the abstract or the explanation with the succinct form to the documents.

Thus, the user does not need to glance over the full text to be possible to understand the documents or the documents set overall content. The text abstract is very useful in some situations, for example, search engine when to user returns inquiry result, usually needs to give the documents the abstract.

D. model quality appraisal

Carries on the excavation in the text to be possible to regard as is one kind of machine learning process. The study result is the knowledge model, carries on the appraisal to the knowledge model is the machine learning important component. The typical assessment method is to the text retrieval basic measure.

{relevant}: With some inquiry related documents set.

{retrieved}: The system retrieves documents set.

{relevant} ∩ {retrieved}: Both are related and the actual documents set which retrieves.

precision: Both are related and the actual documents which retrieves with the documents percentage which retrieves.

recall: Both are related and actual documents which and the inquiry related documents percentage retrieves.

IV. TEXT CLASSIFICATIONS

The text classification is refers to according to the subject category which defines in advance, determines a category for documents set's in each documents. Thus, not only the user can glance over the documents conveniently, moreover may make the documents through the limit hunting zone the search to be easier. At present some websites use the man-power to carry on the classification to the Web documents, some websites use the automatic sorting. The text classification technology algorithm has many kinds, the commonly used algorithm has TFIDF and Nave Bayes and so on.

A Generally method

Will have classified in advance the documents take the training regulations.

Obtains the disaggregated model from the training regulations (to need test procedure, unceasing refinement).

With the disaggregated model which derives to other documents classifies.

B Based on connection taxonomic approach

Proposes the key words and the glossary through the information retrieval technology and the connection parsing technique.

Uses the existing part of speech production key words and the word concept level (documents category).

Discovers the associated word using the connection excavation method, then differentiates each kind of documents (each kind of documents to correspond a group of connection rule).

Goes with the connection rule to the new documents classification.

C Web documents automatic sorting

Uses in the ultra link the information to carry on the classification, the commonly used method includes:

Statistical method

Markov random field (Markov Random Field, MRF)

Unifies loose marking (Relaxation Labeling, RL)

V. TEXT CLUSTERS

The text cluster and the classified difference lies, the cluster has not defined the good subject category in advance, its goal is divides into the documents certain kinds, the request identical kind in documents content similarity is as far as possible big, but the different kind of between similarity is as far as possible small. Hearst et al. the research had already proven “the cluster supposition” the question, namely approaches with the inquiry related documents cluster's comparison, and is far away from the non-correlated documents. Therefore, the documents which will search using the cluster technology divides into certain

kinds. At present has many kinds of text cluster algorithm. Divides into two big types approximately: Level cluster and plane allocation method.

A Level cluster law

Concrete process:

Documents collection D= {d1,…, di,…, dn} each documents di regards as has single member's kind of ci={di}, these kinds constituted D cluster C= {c1,…, ci,…, cn};

Calculates in C every time to the kind (ci, cj) between similarity SIM (ci, cj);

The selection has the biggest similarity kind to arg max SIM (ci, cj), and ci and the cj merge is one new kind ck=ci ∪ cj, thus constitutes D new kind of C= {c1,…, cn-1};

Is redundant the above step, is only left over one kind until C.

Materially this process constructed one to contain in the kind of level information as well as during all kinds and the kind of similarity spanning tree. Because each time merges time, needs overall situation quite all kind of between the similarity, then choice best two kinds, therefore the operating efficiency is not high, does not suit in the massive documents set.

B Plane allocation method

The plane allocation method is documents collection D= {d1,…, di,…, dn} horizontal divides for certain kinds, concrete process:

The determination must produce kind of numberk;

Produces k cluster center according to some kind of principle to take the cluster seed S= {s1,…, sj,…, sk};

To D each documents di, calculates it and each seed sj similarity SIM in turn (di, sj);

The selection has biggest similarity seed arg max SIM (di, sj), belongs to di take sj as cluster center kind of Cj, thus obtains D cluster C= {c1,…,ck};

The redundant step 2~4 certain times, by obtains the stabler cluster result.

This method speed is quick, but k must determine in advance, seed selection difficulty.

the text cluster also has the k-means algorithm, the simple Baye cluster law, the K- most close neighbor to refer to the cluster law, the graduation cluster law as well as based on the concept text cluster and so on.

VI RELATED CONTENTS

Text excavation besides above several introduction content, but also has the following related content research:

Chinese character input and Chinese corpus.

Text phrase delimitation and syntax labelling.

Electronic dictionary construction.

Terminology database.

Machine translation.

Computer auxiliary text proofreading.

Information automatic retrieval system.

Chinese speech recognition system.

Chinese speech synthesis system.

Chinese character recognition system.

The related text excavation's product model has the IBM text intelligence excavator (the hard core is TextMiner, its major function is the feature extraction, the documents accumulation, the documents classification and the retrieval; Supports 16 languages many kinds of form text data retrieval; Uses the deep level the text analysis and the index method; Supports the full-text search and the index search, the search condition may be the natural language and the Boolean logical condition.), the Autonomy Corporation most core's product is Concept Agents (can extract concept automatically from text) as well as Tsinghua University's TH-OCR Chinese character recognition system (recognition precision reaches above 98%).

VIICONCLLUSIONS AND FORECAST

The text excavation, needs to use the natural language processing technology inevitably, constructs the large-scale real text the corpus is the most foundation work. This article elaborated the content is in the text excavation key job. if the foundation work is not solid, the text excavation is very difficult on a big stair. Basic research's foresightedness ought to be able to guarantee in technical the sophistication. Future text excavation technology should be the knowledge retrieval, the knowledge retrieval development should be able the effective addressing following some key questions: (a). Structurized data and non-structurized data mix retrieval; (b) Half structurized content retrieval XML content retrieval; (c).Engine intellectualization knowledge retrieval.

REFERENCE

[1] C.Faloutsos. *Access Methods for Text.* ACM Comput. Surv. , 17 p49-74, 1985.

[2] G.Salton. *Automatic Text Processing*. Reading, MA: Addison-Wesley,1989.

[3] C.J.Van Rijsbergen. *Information Retrieval.* Butterworth,1990.

[4] C.T.Yu and W.Meng. *Principles of Database Query Processing for Advanced Applications*. San Francisco: Morgan Kaufmann,1997.

[5] K.Wang,S.Zhou,and S.C.liew. *Building Hierarchical Classifiers Using Class Proximity.* In Proc.1999 Int. Conf. VLDB'99,P363-374,Edinburgh,UK,Sept.1999.

[6] P.Raghavan. *Information Retrieval Algorithms:A Survey.* In proc.1997 ACM-SIAM Symp. Discrete Algorithms, p11-18,New Orleans,LA,1997.

[7] J.M.Kleinberg and A.Tomkins. *Application of Linear Algebra in Information Retrieval and Hypertext Analysis.* In proc. 18th ACM Symp. Principles of Database Systems,P185-193,Philadelpgia,PA,May 1999

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

A Noval System For Early Detection Of Thyroid With Graph Cluster Ant Colony Optimization

Sayyad Rasheeduddin,
Asst.Professor,
CSE Department,
Malla Reddy College of Engineering

Ch.Vijaya Kumari
Assoc. Professor,
CSE Department,
Malla Reddy College of Engineering

ABSTRACT: Thyroid nodule is defined as an endocrine malignancy that occurs in humans due to abnormal growth of cells. Recently, an increasing level of thyroid incidence has been identified worldwide. Thus, it is necessary to detect the nodules at an early stage. Ultrasonography is an important tool that is utilized for the detection as well as differentiation of malignant thyroid nodules from benign nodules.. Further, large number of features available in US characteristics increases the computation time as well as complexity of classification. In this paper, Graph-Clustering Ant Colony Optimization based Extreme Learning Machine approach is proposed to achieve efficient diagnosis of thyroid nodules. It will enhance thyroid nodule classification by selecting only the optimal features and further using it for improving the function of classifier. The main goal of this technique is to differentiate the malignant nodules from the benign nodules. The performance of both feature selection and classification are evaluated through parameters such as accuracy, AUC, sensitivity and specificity. From the experimental results, it is revealed that the proposed method is significantly better than the existing methods. Thus, it is considered to be an effective tool for diagnosing the thyroid nodules with less complexity and reduced computation time.

Keywords.: Thyroid nodule, ultrasound image, diagnosis, feature extraction, nodules classification.

I. INTRODUCTION

Thyroid is a butterfly shaped small gland situated in the lower region of neck under the layers of skin and muscle. The abnormal growth of cells in the thyroid glands referred to as thyroid nodules. These nodules may be either benign or malignant commonly called as non-cancerous or cancerous cells respectively (Acharya et al, 2016). Thyroid nodules are the most common search criteria in thyroid gland as it is present in almost 40% of the population among world-wide and about 5-10% is found to be malignant. Thus, radiologists are involved in diagnosing the thyroid gland to identify the risk of malignancy with respect to the guidelines provided by Thyroid Imaging, Reporting and Data System (TI-RADS). In general, thyroid nodules affect both men and women whereas it is severe in case of women and its formation depends upon different characteristics like gender, age and population (Erdem et al, 2010). The thyroid disease analytics have revealed that thyroid is a severe disorder which increases the mortality rate in humans.

Therefore, it is necessary to produce an accurate tool for malignancy risk detection in order to increase the survival rate of thyroid patients. Moreover, early identification of the symptoms of thyroid disorders can improve the survival rate thereby initiating the treatment at initial stage (Koundal et al, 2018). However, the diagnosing process as well as treatment of thyroid disease remains difficult and the main challenge in this field is differentiation between the nodules. It is necessary to accurately classify the thyroid nodule because of high prevalence of the nodules as well as less prevalence of the malignancy.

Fine needle aspiration biopsy (FNAB) is the standard treatment utilized for diagnosing the thyroid diseases but it is reported that it can mimic other kind of diseases (Bakshi et al, 2003).Several thyroid treatment plans use FNABs as reference since it is labor-dependent and expensive under large scale diagnosis. Likewise, unwanted biopsies cause anxiety, irritation and increase the treatment-expense to thyroid patients (Ma et al, 2017). Even though massive growth is achieved in the field of thyroid diagnostics with sources such as CT imaging, radionuclide and MRIs still it is necessary to select an appropriate and stable material for effective differentiation between the nodules (Wu et al, 2013). Furthermore, clinical procedures do not obtain better diagnosis and so, non-invasive imagery tools like Ultrasonography is identified as a best choice for distinguishing among the nodules. The American Thyroid Association (ATA) stated that ultrasound images are the primary choice of any radiologists for examining the thyroid nodules. Furthermore, if a nodule is identified on other kind of image modalities, detailed diagnostics are performed on US images. Thus, Ultrasonography is the initial stage diagnosing modality for thyroid disease identification. (Cooper et al, 2006) defined that US images are sensitive and suitable for examining the nodularity of thyroid compared to other images such as MRI and CT. Sonography visualize the different characteristics of thyroid glands like dimension, structure, echogenicity, availability of calcification, etc. In literature, more number of research works has been carried out to distinguish between benign and malignant nodules as it is necessary to provide proper and effective treatment to thyroid patients. Thyroid nodules are comprised of different kind of textural features. Ultrasound images also resemble numerous features like electrographic, textural and morphologic which are important for the purpose of nodule classification. In the last decade, medical assistants experimented different sonographic features to prove its efficiency in diagnosing the risk of malignancy of thyroid disorders. However, feature selection is the main task of many machine learning based disease classification techniques Feature Selection 2010). A number of approaches have been developed to extract relevant features from US images. It is identified that the process of feature extraction and using them to train a classifier consumes more amount of time. Therefore, certain features are neglected during diagnosis inorder to reduce time and to improve the reliability of classification. A

proper methodis required to select appropriate features and to neglect irrelevant ones. Classification and prediction achieves accurate result with limited features than processing with all the available features. Thus, proper feature extraction and classification techniques are needed to attain better results in disease prediction. (Tsantis et al, 2009) presented CAD based diagnosis system that used morphologic and wavelet-based features for classifying the thyroid nodules in US images. These features are extracted based on malignancy related characteristics like calcification, irregular shape, uniformity, echogenicity, etc. The efficiency of using extracted features in classification is evaluated through two different pattern recognition algorithms such as probabilistic neural network and support vector machine. It showed that the extracted features can improve the accuracy of classifier and lowers the faults in thyroid disease identification.

In general, computation time and prediction accuracy are the important aspects that are taken into consideration during thyroid nodules differentiation. Therefore, in this research, an improved fast learning based pattern recognition tool called as ELM is utilized for the prediction of thyroid disease with US characteristics. Extreme Learning Machine (ELM) is a new learning based approach that supports single hidden layer feed-forward neural networks (SLFNs). Compared to gradient-based methodologies that iteratively adjust the parameters of neural network, ELM randomly selects the input weights as well as hidden biases for the determination of output weights by adopting the generalized inverse of Moore–Penrose (MP) analytical method. Further, it learns faster with highly generalized performance and also keeps the parameter tuning-free. Due to these properties, ELMs are widely used on classification areas like predicting patient outcomes (Liu et al, 2011), sales forecasting (Chen & Ou, 2011) and so on. Moreover, ELM proved its reliability on number of disease diagnosing tasks over other learning based classification algorithms.

In this paper, a machine learning model is proposed to achieve efficient classification of the thyroid nodules. Existing classification techniques have recommended that using an optimal feature selection process will enhance the accuracy of classifier used for nodule differentiation task. Hence, anovel feature selection method called as graph-clustering based ant colony optimization is adopted in this proposed work. It will select the discriminant features thereby making it effective for classifying the nodules within limited time.Further, Extreme learning machine based classifier differentiates the benign and malignant nodules. The maincontribution of this paper is defined as follows:

- A hybrid methodology is proposed to enhance the diagnostics of thyroid disease using ultrasound characteristics
- An optimal feature selection approach called as graph clustering based ant colony optimization tool is applied to extract the relevant features from the raw dataset
- To increase the efficiency as well as accuracy in differentiation of thyroid nodules, a computer aided diagnosis system based on extreme learning machine is also proposed. The remainder of this paper is organized as follows. Section 2 provides a literature review on different feature selection and machine learning based classification approaches. Section 3 presents a background review on graph clustering based ant colony optimization and extreme learning machine classifier. The detailed implementation of proposed methodology is given under Section 4. The experimental results and discussion of the proposed method is visualized in Section 5. Finally, the conclusion of this paper is defined in Section 6.

II. LITERATURE REVIEW

Recently, a number of thyroid disease diagnostic systems were introduced to analyze the severity of thyroid disorders using ultrasound characteristics. It includes computer aided diagnosis (CAD) (Sollini et al, 2018), deep convolutional neural networks (Li et al, 2019; Li et al, 2019), machine learning and so on. As CADs works on the principle of machine learning algorithms, it is mostly preferred by radiologists for identifying the risk of malignancy in thyroid glands. It is identified that machine learning based thyroid disease diagnostic systems would increase the accuracy of analysis with ultrasound imaging.(Ardakani et al, 2015) identified a new approach to analyze the texture of US images based on computer aided diagnosis (CAD) to distinguish the thyroid nodule as benign or malignant. The Receiver Operating characteristic Curve (ROC) analysis showed that Texture Analysis (TA) is a reliable approach which provides useful information to identify and to classify the nodules. Furthermore, this technique consumes low cost and does not need any human intervention as the entire diagnosis is performed on computers. But is tested on small datasets along with highly sensitive FNAB approach. FNABs need the help of surgical pathology to obtain more definitive results and it excludes certain data due to an indeterminate operation.

In classification based applications, features are the important factors that impact the discriminatory functioning of the classifier. Generally, it is effective to consider all the features during classification but it is redundant due to mutual correlation between them. Due to this, it is enough to select the relevant features from the available dataset which will then increase the classification accuracy of classifiers. However, the main task is to choose the suitable feature selection approach for the particular classification algorithm. On the other hand, selecting the optimal feature selection method can improve the classification accuracy but increases the time as well as computation complexity. In the last decade, number of feature selection algorithms was introduced that include random searches, heuristic, greedy and exhaustive. However, these techniques are computationally very expensive and get trapped into local optima. To overcome such situations, different kind of feature selection methods like Ant Colony Optimization (ACO) (Tabakhi et al, 2014), Genetic Algorithm (GA) (Kabir et al, 2011) and Particle Swarm Optimization (PSO) (Yong et al, 2016) are presented. Of these, ACO seems to be very affective as it is multi-agent based selection methodology. The advantages of ACO over swarm intelligence based techniques include its local and global optima ability, availability of long-term distributed storage, and utility of reinforcement based machine learning concept. (Choi et al, 2015) provided a systematic approach to quantitatively observe the features of US images on calcified

pathological thyroid nodule dataset. The features responsible for tumor malignancy were identified by means of a univariate algorithm and the nodules are differentiated by using neural network approach. The diagnostic ability of both the neural network and feature estimation algorithm sarederived from ROCs and AUCs (Area Under ROCs) respectively. However, the application of this technique is limited to descriptive 2D-calcified datasets as it is necessary to perform retrospective review on those datasets. In addition, only the dataset with surgical cases are taken for analysis and it does not consider clinically visible benign nodule dataset. Conversely, existing CADs cannot provide promising results and several radiologists have reported that its clinical usage is limited on certain practices.

(Ouyang et al, 2019) estimated the performance of linear as well as non-linear machine learning approaches in thyroid nodule malignancy identification with reference to a standard approach. The diagnostic performance analyzed through AUC showed similar AUCs on non-linear techniquescompared to linear techniques. Particularly, Kernel SVM and Random Forest algorithms attained moderately larger AUCs compared to other algorithms taken for observation. As these analyses does not perform any pre- processing or feature extraction tasks on image datasets, it is found to be an enhanced approach than the CAD system. The echographic view of thyroid nodule in thyroid imaging is referred to as texture. In mathematical model based diagnosing applications, these textures are analyzed by means of quantitative parameters. This is very helpful for CAD based disease diagnosing applications.

The heterogeneous nature of thyroid nodules, presence of different internal substances and large number of echo patterns in US images confuse the physicians and radiologists to identify the appropriate textures. Thus, textural feature extraction techniques are introduced to distinguish suitable texture patterns thereby decreasing the misdiagnosis rate. (Chang et al, 2010) tested six kinds of SVMs in using important textures and to increase the efficacy of thyroid lesion classification. The experimental outcomes proved the reliability of their method in extracting the important features from thyroid imagery. It is then compared with an existing approach called as sequential- floating-forward-selection (SFFS). This comparison showed that the performance of SVMs in feature extraction is similar to SFFS but the execution time is 3-37 times faster than SFFS.

(Shankar et al, 2018)established a kernel-based classification model to classify thethyroid nodules after selecting appropriate features from thyroid dataset. Grey wolf optimization based feature selection algorithm is adopted to improve the dataset classification. Their technique showed improved performance on dataset classification but it consumes large amount of time. (Han et al, 2006) discovered an ELM model to predict how long a non-small cell lung cancer postoperative patient can survive. This method showed accurate results in prediction and the convergence rate is faster than the ANN framework. (Zhang et al, 2007) determined the functioning of ELM model on multi-categorical classification of microarray dataset of cancer patients. It was observed that the classification accuracy, training time and the computation complexity of ELM are better than ANN and SVM classifiers. (Helmy&Rasheed, 2009) utilized ELMs to diagnose five types of diseases, and the classification accuracy and computational complexity for this observationis effective with reduced training dataset. (Gomathi & Thangaraj, 2010) suggested computer aided ELM lung cancer diagnostic system. The experimental outcome of this system is compared with SVM approach where ELM produce more accurate results in classification task.

(Li et al, 2012) recommended computer operated diagnostic system that works on the basis of principal component analysis and extreme learning machine. It performed thyroid diagnosis by taking into account different characteristics of ELM like simplicity, less complexity, generalized behavior, faster learning capability and less time for computation. Further, feature extraction is performed through PCA which ignores irrelevant data and consider only the appropriate features for ELM classification. This technique is efficient in classifying three different forms of thyroid disorders like hyperthyroid, hypothyroid and normal thyroid. PCA-ELM classifier is precise and it provides accurate categorization of thyroid malignancy. However, still is a challenging task to provide timely efficient feature selection approach for selecting appropriate features and using it for differentiating the nodules. Further, existing classifiers are highly complex which performs larger calculation throughout the entire operation. To overcome these issues, a hybrid approach is introduced in this paper by combining graph-clustering ant colony optimized feature extraction with extreme learning machine classifier.

III. BACKGROUND METHODOLOGY

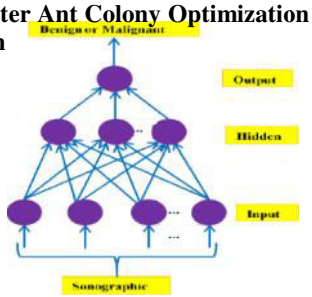
A *Ant Colony Optimization*

(Sivagaminathan & Ramakrishnan, 2007) explained that ant colony optimization is similar to the real-life behaviour of ants while travelling through the same path to reach their nest after collecting the food. It should be noted that it does not make any visual contact with the path itis travelling. This could be achieved through an indirect mode of communication known as “stigmery” by an odorous chemical substance named as pheromone. The quantity of the pheromone substance depends upon different properties of the food source like quantity, quality and distance of availability. All the ants follow a path that contains more pheromone thereby making it a positive feedback loop. The pheromone starts vanishing and evaporating after certain time period, and finally result in reduced pheromone in less demand path. Because of pheromone evaporation, the ants search other available paths and finalize the most optimal path for travelling. By following the same procedure, the optimal features are selected from the thyroid dataset using ant colony optimization algorithm. Due to its simplicity, it is found to be a better method for machine learning based classification designs.Nevertheless, the use of present ACO algorithms in recent application suffer from several drawbacks like computation complexity due to fully connected graph structure drawn from all the available features, need for a learning model to create feature subsets, highly correlated

Extreme

ISBN 978-93-88808-61-3

with Graph Cluster Ant Colony Optimization Based Feature Selection



To address the above issues, (Moradi&Rostami, 2015) introduced a new feature selection approach by integrating ant colony optimization with graph-clustering technique. Graph Clustering based Ant Colony Optimization is a filter based multivariate feature selection process that represent the features in the form of undirected graph with nodes and edges of the graph as features and similarities between features respectively. Similar to data clustering algorithms like Fuzzy C Means (FCM) and k-means, graph clustering initially identifies the similarity among pair of data points and form an undirected graph. This graph is then divided into clusters according to the linear/non-linear boundaries estimated through an optimal objective function. In GCACO algorithm, feature clustering is performed by means of a

community detection algorithm, which produces a subset of features with each feature containing minimal redundancy with other features available in the subset. The key role of this technique over other techniques such as F-Score, L- Score, ReliefF and UFSACO is that it considers both redundancy and relevance analyses while performing the feature selection task. Likewise, GCACO identify the relationship between features before performing selection whereas univariate approaches rank the features without considering the dependency between each feature. Since graph clustering finds advantages by integrating with ant colony optimization, it is extensively used for feature selection in many applications.Graph Clustering based Ant Colony Optimization feature selection approach is adopted in the proposed work for selecting necessary features from the thyroid dataset. This is an effective approach to select discriminant features that are very essential to differentiate the thyroid nodules from the dataset.

B. Extreme Learning Machine

The use of extreme learning machine (ELM) as novel machine learning algorithm for single layer feed forward neural networks (SLFNs) displayed in Fig. 1 was first initiated by Huang et al (2004). It overcomes the drawbacks of conventional SLFNs related to slow learning speed, tuning of trivial components and improper generalization ability. Therefore, ELM possesses different properties such as fast learning capability, highly generalized performance and free parameter tuning.ELM is designed in such a manner to function well with enhanced generalization capability for performing better classification and Fig 1. Structure of Extreme Learning Machine

IV. PROPOSED METHOD

The proposed method is designed to predict the thyroid disease from dataset by classifying the thyroid nodules using US features. Firstly, the discriminate features are partitioned from the dataset using Graph Clustering based Ant Colony Optimization feature selection method. Secondly, each of the selected features is experimented to differentiate the type of nodule using Extreme Learning Machine algorithm.

A.Feature Extraction using Graph Cluster Ant Colony Optimization

GCACO is a multivariate feature selection strategy that selects the optimal features by performing dependency analysis on features structured as an undirected graph**.** In order to select the discriminant features, the relevance analysis is performed by means of the Fisher score (F-Score) and the multiple discriminant analysis (MDA).

i. Graph formation. Initially, the features are analyzed one- by-one to know about its redundancy. This analysis is achieved by creating a weighted undirected graph with all the available features. The graph representation is defined as: , where depicts the nodes of the graph, represents the edges between the graph nodes, and denotes the weight of edges of the graph. The weight between nodes and is calculated as: regression. On comparing the learning processes of both ELM as well as SLFNs gradient based iterative and back propagation methods, ELM learns faster than SLFNs.

Where, and are the features and and are the mean value of the feature vectors. The feature vectors and is either extremely correlated or uncorrelated when the weight of their edges produce a value one or zero respectively. The weight value in GCACO is normalized by means of softmax scaling technique inorder to get rid of the impact of outliers. The procedure of softmax scaling is defined as follows:

Where and are the mean and standard deviation of all values of .

A. Feature clustering. Redundancy analysis in GCACO is performed by an effective algorithm called as Louvain community detection. In this, the weighted undirected graph is partitioned into communities or sub-nodes depending upon the similarities between features (highly correlated). During initialization, each node (feature) is treated as an individual community. For each iteration, two neighbors (say and) of a node is chosen. A modular gain factor is evaluated by eliminating the node from its own community and inserting it to one of the communities of node . This procedure is continued until all the Discriminant Analysis (MDA) and the final subset of features is sorted according to their pheromone value. The implementation process of GCACO algorithm is explained below:

Step1:The algorithm is initialized by setting up the following parameters: total iterations , number of ants , evaporation coefficient of pheromone , initial pheromone quantity and other constants ,□, .

Step2: In GCACO, the relevant feature analysis isdone by the Fisher score value. Based on the F score value, the features are sorted according to their significance. The Fisher score for the th feature in the feature set is defined as:

neighbors of node are visited. The node is then added into the community that results in higher positive modular ain factor. If the modular gain of all the neighboring communities is negative, then remain in its own community. This procedure is repeated until no change is found in the modular gain value and a new network is drawn based upon the final communities. The modularity gain factor that is obtained after inserting the isolated node to any of one of the communities is determined as: Where and are mean and standard deviation of the th class with samples respectively, and is the mean of the samples in the th feature vector. The F score value is normalized within range 0 and 1 by the softmax scaling method. The features ith largest F-score value are considered to be better discriminate features.

Step 3: In GCACO, the redundancy analysis is achieved by calculating the absolute value of Pearson's correlation. To obtain this, the cross-correlation mean values within th feature and all other features visited by the th ant from all previous clusters is first evaluated. Finally, the following function is estimated to know about the redundancy:

Where is the total weight of interior edges of community , is the total weight of edges that are incident to nodes in community , is the total weight of edges that are connected to node is the total weight of edges from node to every node in community, andThis function is utilized to know about exploitation/exploration of feature during feature selection task. In this function, represents the size of .To enhance the outcomes of feature , then the th ant choose the succeeding feature as

clustering, the weights less than , a preset threshold value used to control the amount of clusters are excluded, where ranges between 0.3 and 0.8.

ii.Ant Colony Optimization. The working of Ant Colony Optimization algorithmis based on the movement of ants through the path travelled by other ants identified by pheromone chemical imprintsleft by them.In GCACO, anant is randomly allocated to one of the clusters produced by feature clustering approach. On each iteration, the algorithm considers two random values and along with a parameter ( ) and threshold value (□). The value of , , and □ lies between therange 0 and 1. If , the ant selects a feature from the cluster on the basis of roulette wheel concept. If □, the ant remains in the same cluster follows:

Where represents the features that are still not visited by the th ant from the present cluster (th cluster), is the total pheromone of the th feature, and and are the relative importance of the pheromone value and heuristic information, respectively.

(ii) If , a probability function is estimated for the remaining features in the present cluster. It is described as follows:

and choose another feature. When □, the ant leaves the current cluster and goes to another cluster. The parameter is used to switch among exploitation and exploration phase and the threshold value □is utilized to control the number of features to be selected within a cluster. The above process is continued until selecting features from all the available clusters. After going through all the clusters, the features selected by first ant are stored and the next ant enters into feature selection process. The same procedure is repeated in a cyclic order for the required amount of iterations. The pheromone values of the features are maintained by Multiple Then, the next suitable feature is chosen on the basis of roulette wheel rule.

Step 4:On every iteration , the pheromone amount of the th feature is updated based on MDA as follows: As explained by Huang et al. [50], the output matrix of the hidden layer of neural network with th column of being the th output of hidden neuron with respect to the input variables, Further, they showed that the hidden layer bias and input weights of SLFNS are not expected to be modified and are provided in a random

Where represents the feature selected on th iteration by the th ant, denotes the number of ants and is defined manner. With this assumption, the output weights are analytically estimated through least square solution of the as the separability indexof th selected subset in the thlinear system, :

iteration, i.e., and are the between and within scatter matrices respectively, is the transform matrix from the

-dimentionalspace to the -dimentional space, where is the number of the features selected by the th ant in the th iteration and is an integer value between 1 to with as the total number of classes.

Step 5:After completing the overall iterations(, the value of pheromone is utilized to select the needed optimal feature set. In each cluster, the features are sorted according to the amount of pheromone content and the first set of features from each cluster is chosen for further processing. Therefore, for clusters, features are selected.

B. Extreme Learning Machine based classification

A brief description about classification using Extreme Learning Machine is explained in this section. Consider a training set where and denotes the input feature vector of size and target vector of size respectively. The conventional SLFNs hold an activation function and the amount of hidden neurons can be mathematically framed as follows:

Where and represent the weight vectors among the input layer and output layer of th neuron in the hidden layer respectively. is the bias of the th neuron in the hidden layer and is the target vector of the th input data. The inner product of and be . If it is possible for the SLFNs to approximate the samples with zero errors, there will be i.e., and there exists , , such that , = , ,..., . The above Equation can be reformulated as follows:

Where, min The above Equation can be easily determined by a generalized linear approach like Moor-Penrose (MP) by finding the inverse of , as is shown inthe Equation given below.

Where is the generalized inverse matrix obtained from MP approach. Utilizing this generalized inverse may result in minimized solution for the resulting least square norms. It yields the unique as well as smallest least square norms compared to existingleast square solutions. After performing effective analysis, Huanget al. [49] explained that the generalized inverse of MP obtains better ELM performance with dramatically improved learning speed.The learning process of ELM is proceeded as follows: Initially, consider a training set, , , an activation element , and total hidden neurons . (a) Randomly allocate the input weights and bias (b) Evaluate the output of Hidden layer matrix . (c) Estimate the resultant weight .

V. RESULTS AND DISCUSSION

ISBN 978-93-88808-61-3

The implementation and performance analysis of the proposed work is performed on MATLAB R2018a software running on windows operating system with 1.7 GHz CPU and 4.00 GB of RAM. This analysis is performed to know about the functioning of feature selection as well as classification approaches. The extreme learning machine is built on the basis of 10-fold cross validation process on the thyroid disease dataset.

Experimental Design

A. Feature selection

The retrospective analysis of thyroid disease is performed on pathologically verified thyroid nodules with the help of different characteristics of US images. The thyroid dataset taken for evaluation consists of1427 nodules with 1180 benign nodules and 247 malignant nodules. The benign nodules are considerably lengthier than the malignant nodules. For evaluating the performance of feature selection algorithm, different characteristics of US images were considered. It includes different features like demographic information, boundary, echo pattern, posterior acoustic pattern, margin, orientation, position, thyroid shape, tumor and size and calcification. These features are extracted from the US images. After applying GCACO algorithm, the important features are selected from the overall available features as shown in Figure 2 and Table 1.

The definitions o selected features are provided below.

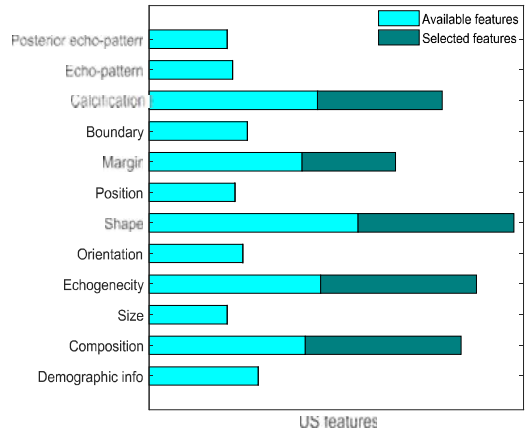


Fig 2. Features selected from the thyroid dataset by using GCACO approach

Table 1. Feature subset selected by GCACO algorithm

Size of subset	Features selected on each iteration
1	{Calcification}
2	{Calcification, Composition}
3	{Calcification, Composition, Echogenicity}
4	{Calcification, Composition, Echogenicity, Shape}
5	{Calcification, Composition, Echogenicity, Shape, Margin}

- **Calcification.**Calcification is categorized into three types like microcalcification, macrocalcification or no calcification. Microcalcification and macrocalcification are calcification with diameter less than 1 mm and larger than 1 mm respectively. If a nodule consists of both types of these calcifications, then it is remarked as microcalcification.
- **Composition.** The proportion of fluid or soft tissue in a nodule is termed as composition. It may be solid or liquid or cystic. Solid is comprised of soft tissues with liquid lesser than 10%. Predominantly solid substances are consisted of >10% liquid on<50%volume of the nodule. Incase of cystic composition, the nodule is fully or almost fully filled with liquid. One special appearance of composition is spongiform appearance that resembles like minute cystic spaces detached by thin pieces of septa.
- **Echogenicity.** In solid portions, the echogenicity is classified as iso/hyper-echogenicity, hypo-echogenicity and marked hypo-echogenicity. If the echogenicity in the nodule looks similar to the thyroid parenchyma present in their surroundings, it is termed as iso-echogenicity. If the echogenicity is low as compared to that found in strap muscles, it is called as marked hypo-echogenicity.
- **Shape.** The shape of the thyroid gland may be oval or round and it is either taller than wide or taller than long. The shape of the nodules is identified from the diameter of an anteroposterior nodule. If the anteroposterior diameter is smaller than the diameter of longitudinal and transverse planes, then the shape is said to be oval shape. Otherwise, if the anteroposterior diameter is equal to the diameter of longitudinal and transverse planes, then it is called as round shape. If the ratios of anteroposterior to transverse and longitudinal diameters are greater than one, then the structure of nodule is taller than wide and taller than long respectively.
- **Margin.** The outline of thyroid nodule is called as margin of the nodule. The margin of the nodule takes different structures like smooth margin, ill-defined margin, irregular margin and microlobulated.

Table 2. Ultrasound features of thyroid nodules

Features	Number of benign nodules (n=1180)	Number of malignant nodules (n=247)	*p*-value (analysis)
Calcification			<0.001
Macro-calcification (n=303)	257	46	
Micro-calcification (n=127)	50	77	
No calcification (n=997)	873	124	
Composition			<0.001
Solid (n=983)	763	220	
Mixed (n=444)	417	27	

Echogenicity			<0.001
Hyper-echogenicity (n=854)	650	204	
Hypo-echogenicity (n=412)	370	42	
Marked hypo-echogenicity (n=161)	160	1	
Margins			<0.001
Smooth (n=1010)	985	25	
Microlobulated (n=243)	178	65	
Irregular (n=174)	17	157	
Shape			<0.001
Wider than tall (n=1253)	1117	136	
Taller than wide (n=174)	63	111	

The information about the extracted features is given in Table 2.These features are significant to identify the risk of malignancy associated with thyroid nodules. From Chi- square analysis, the nodules with the following criteria are considered as malignant

Extreme Learning Machine for Thyroid Nodule Classification with Graph Cluster Ant Colony Optimization Based Feature Selection

Table 3. General structure of the confusion matrix

Type of classifier		**Predicted**	
		Malignant nodules	**Benign nodules**
Actual	Malignant nodules	True positive	False Negative
	Benign nodules	False Positive	True Negative

The equations for calculating accuracy, sensitivity and specificity from the confusion matrix are described above. The region under the receiver operating characteristic curve is termed as the area under curve which draws the true positives rather than the false positive rates. Finally, the classifier with higher AUC is remarked as the best classifier than the classifier that produces smaller AUCs. A classifier with AUC equal to one is concluded as a perfect classifier.

B. Performance Analysis

The performance of the proposed thyroid disease nodule classification method is evaluated by analyzing the functions of both GCACO feature selection and extreme machine learning classification algorithms. The effectiveness of classification based on the selected features is compared with existing methods such as L-Score, F-Score, ReliefF and UFSACO. Table 4-6 shows the comparison table for classification accuracy, sensitivity and specificity obtained with different feature selection algorithms. Different feature criteria that decides the efficiency of classification includes composition, calcification, margin, shape, solid part of echogenicity and size of nodules that are equal to or larger than 5 mm.

The comparison of classification accuracy of GCACO with existing feature extraction methods such as L-score, F-score, ReliefF and UFSACO are shown in Table 4. From this analysis, it is clear that GCACO produce better accuracy in classification than the existing methods. The maximum classification accuracy obtained for GCACO is 98%, 95% and 97% with feature subsets 1, 4 and 5 respectively. The featuresubsets extracted from relief methods also produced better results with 96% and 95% accuracy for number of features 2 and 3 respectively. On the other hand, L-score, F- score and UFSACO produced worst results on classification accuracy. The classification sensitivity for GCACO outperformed the existing methods with 98%, 96% and 94% sensitivity with subsets 1, 3 and 4 respectively. It is listed in Table 5. The ReliefF method utilized 2 and 5 number of featuresfor performing the classification. The sensitivity of these featuresubsets are found to be 97% and 95% respectively. As shown in Table 6, the classification specificity of GCACO for the subsets 1, 2, 3 and 4 is found to be 99%, 95%, 98% and 95%

Table5. Comparison of classification sensitivity with different feature selection algorithms

No. of features in feature subset	**L-score**	**F-score**	**ReliefF**	**UFSACO**	**Proposed**
1	86.11±6.56	89.35±1.52	94.67±2.77	95.35±3.45	**98.56±1.24**
2	89.46±4.38	95.67±4.45	**97.46±1.56**	86.56±2.35	93.67±4.16
3	84.46±2.45	90.78±7.54	90.67±3.86	88.67±4.26	**96.67±4.24**
4	94.45±6.35	94.24±7.87	91.35±7.45	92.57±4.26	**94.46±1.23**
5	95.47±3.88	95.24±1.48	**95.67±2.56**	94.67±3.54	94.57±2.65

respectively. But, L-score produced 91% better classification specificity with 4 features in the feature subset. Thus, compared to univariate feature selection approaches like L-score, F-score and Relief F, the proposed multivariate feature selectionapproach functions significantly better. The performance of proposed ELM classifier is evaluated by comparing it with other classification techniques such as SVM and KNN. The average accuracy, sensitivity and specificity obtained for all these techniques for the features of

thyroid datasets are presented in Table 5. From there, it is observed that the ELM classifier outperforms the existing approaches used for classifying the thyroid dataset with selected feature subsets. Further, the improved accuracy, sensitivity and specificity resemble the higher grade of relevancy and redundancy between the selected features from the input dataset.Thus the significant performance obtained from the ELM classifier is due to high relevancy as well as low redundancy among the features considered for classification. The confusion matrix outcomes of three classifiers such as ELM, ANN and SVM are presented in Table 7. It is identified that ELM has properly differentiated 57 malignant nodules and 87 benign nodules. Further, it misidentified 16 malignant nodules as benign and 10 benign nodules as malignant. On the other hand, ANN perfectly classified 49 malignant nodules and 23 benign nodules. In addition, it misjudges 23 malignant nodules as benign and 17 benign nodules as malignant. Finally, the SVM classifier predicted 51 malignant nodules and 83 benign nodules. However, it misjudges 20 malignant nodules as benign and 14 benign nodules as malignant. This visualize that ELM outperforms the other two methods in differentiating the type of nodules.

Table 7. Confusion matrix of ELM

	Type of classifier	Predicted nodules	
		Malignant	Benign
Actual nodules	**ELM classifier**		
	Malignant	57	16
	Benign	10	87
	ANN classifier		
	Malignant	49	23
	Benign	17	80
	SVM classifier		
	Malignant	51	20
	Benign	14	83

The performance of ELM based on different hidden neurons is shown in Figure 3. As shown, the accuracy and specificity fluctuates through certain limit and remains stable as the total amount of neurons is increased. The sensitivity and AUC is found to be almost stable throughout the hidden neurons. At 20 neuron case, the accuracy of the extreme learning machine classifier is found to be higher and so, it is taken as the optimal number of the neurons.

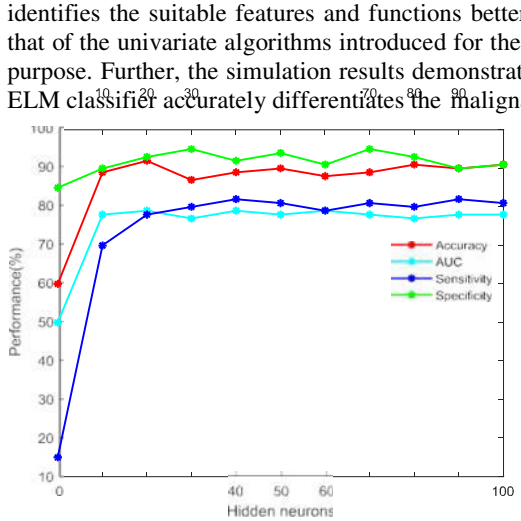


Figure 3. Performance of ELM based on the hidden neurons

The comparative results for accuracy, AUC, sensitivity and specificity of ELM and existing techniques like SVM and ANN classifiers are shown in Figure 4.It is provided in terms of the mean value of different parameters taken for analysis. The figure depicts that the performance of ELM is better than other machine learning classification techniques such as SVM and ANN. Further, the performance of SVM is slightly higher than that of the artificial neural network classifier. This describes that that ELM is the best method to be used for analyzing the thyroid disease using US characteristics.

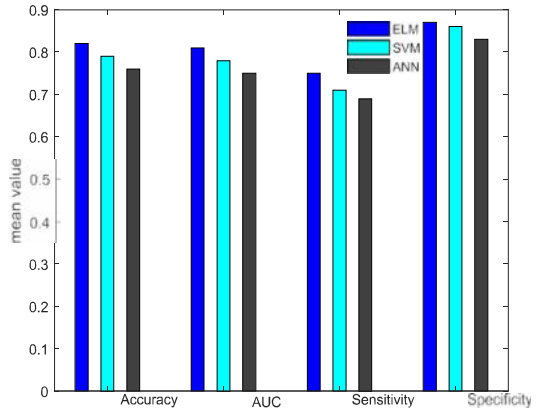


Figure 4. Comparative analysis of classifier algorithms in terms of: accuracy, AUC, sensitivity and specificity

VI. CONCLUSION

The ultrasound is defined as a non-invasive tool that is utilized for the diagnosis of thyroid lesions because of its affordable price and ease of availability. During thyroid treatment, differentiation of malignancy nodules remain a critical task due to the challenges faced by texture analysis and employing machine learning algorithms in modern diagnostic procedures. Thus, it is necessary to modernize and authorize these processes for the purpose of making it to be desire in the field of thyroid diagnostics. In this paper, Graph- Clustering Ant Colony Optimization based Extreme Machine Learning approach is introduced for the detection of malignancy risk associated with thyroid nodules. Using GCACO, the features are partitioned into group of clusters and the clusters are assembled as an undirected graph with community detection algorithm. Afterwards, ACO is employed to select the optimal features from the group of clusters. In this work, five of the ultra-sonographic features are taken as discriminant features from the US thyroid dataset. The GCACO feature selection method is multivariate and it is compared with existing univariate methodologies like L-score, F-score, ReliefF and UFSACO. It is identified that GCACO significantly identifies the suitable features and functions better than that of the univariate algorithms introduced for the same purpose. Further, the simulation results demonstrate that ELM classifier accurately differentiates the malignant

nodules from benign nodules. Thus, GCACO based ELM classifier can be efficiently applied for clinical diagnosis of thyroid disorders and produce effective result in thyroid treatment.

REFERENCES

1. Acharya, U.R., Chowriappa, P., Fujita, H., Bhat, S., Dua, S., Koh, J.E., Eugene, L.W.J., Kongmebhol, P. and Ng, K.H., 2016. Thyroid lesion classification in 242 patient population using Gabor transform features from high resolution ultrasound images. Knowledge-Based Systems, 107, pp.235-245.
2. Ardakani, A.A., Gharbali, A. and Mohammadi, A., 2015. Application of texture analysis method for classification of benign and malignant thyroid nodules in ultrasound images. Iranian journal of cancer prevention, 8(2), p.116.
3. Bakshi, N.A., Mansoor, I. and Jones, B.A., 2003. Analysis of inconclusive fine-needle aspiration of thyroid follicular lesions. Endocrine pathology, 14(2), pp.167-175.
4. Chang, C.Y., Chen, S.J. and Tsai, M.F., 2010. Application of support- vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images. Pattern recognition, 43(10), pp.3494-3506.
5. Chen, F.L. and Ou, T.Y., 2011. Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry. Expert Systems with Applications, 38(3), pp.1336-1345.
6. Choi, W.J., Park, J.S., Kim, K.G., Kim, S.Y., Koo, H.R. and Lee, Y.J., 2015. Computerized analysis of calcification of thyroid nodules as visualized by ultrasonography. European journal of radiology, 84(10), pp.1949-1953.
7. Cooper, D.S., Doherty, G.M., Haugen, B.R., Kloos, R.T., Lee, S.L., Mandel, S.J., Mazzaferri, E.L., McIver, B., Sherman, S.I. and Tuttle, R.M., 2006. Management guidelines for patients with thyroid nodules and differentiated thyroid cancer: The American Thyroid Association Guidelines Taskforce. Thyroid, 16(2), pp.109-142.
8. Erdem, G., Erdem, T., Muammer, H., Mutlu, D.Y., Fırat, A.K., Sahin,
   I. and Alkan, A., 2010. Diffusion- weighted images differentiate benign from malignant thyroid nodules. Journal of Magnetic Resonance Imaging, 31(1), pp.94-100.
9. Frates, M.C., Benson, C.B., Doubilet, P.M., Kunreuther, E., Contreras, M., Cibas, E.S., Orcutt, J., Moore Jr, F.D., Larsen, P.R., Marqusee, E. and Alexander, E.K., 2006. Prevalence and distribution of carcinoma in patients with solitary and multiple thyroid nodules on sonography. The Journal of Clinical Endocrinology & Metabolism, 91(9), pp.3411-3417.
10. Gomathi, M. and Thangaraj, P., 2010. A computer aided diagnosis system for lung cancer detection using support vector machine. American Journal of Applied Sciences, 7(12), p.1532.
11. Han, F., Huang, D.S., Zhu, Z.H. and Rong, T.H., 2006. The forecast of the postoperative survival time of patients suffered from non-small cell lung cancer based on PCA and extreme learning machine. International journal of neural systems, 16(01), pp.39-46.
12. Helmy, T. and Rasheed, Z., 2009, May. Multi-category bioinformatics dataset classification using extreme learning machine. In 2009 IEEE Congress on Evolutionary Computation (pp. 3234-3240). IEEE.
13. Huang, G.B., Zhu, Q.Y. and Siew, C.K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks. Neural networks, 2, pp.985-990.
14. Kabir, M.M., Shahjahan, M. and Murase, K., 2011. A new local search based hybrid genetic algorithm for feature selection. Neurocomputing, 74(17), pp.2914-2928.
15. Li, L.N., Ouyang, J.H., Chen, H.L. and Liu, D.Y., 2012. A computer aided diagnosis system for thyroid disease using extreme learning machine. Journal of medical systems, 36(5), pp.3327-3337.
16. Li, X., Zhang, S., Zhang, Q., Wei, X., Pan, Y., Zhao, J., Xin, X., Qin, C., Wang, X., Li, J. and Yang, F., 2019. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. The Lancet Oncology, 20(2), pp.193-201.
17. Li, X., Zhang, S., Zhang, Q., Wei, X., Gao, M., Zhang, W. and Chen, K., 2019. Deep convolutional neural network models for the diagnosis of thyroid cancer–Authors' reply. The Lancet Oncology, 20(3), p.e131.
18. Liu, N., Lin, Z., Koh, Z., Huang, G.B., Ser, W. and Ong, M.E.H., 2011. Patient outcome prediction with heart rate variability and vital signs. Journal of Signal Processing Systems, 64(2), pp.265-278.
19. Koundal, D., Gupta, S. and Singh, S., 2018. Computer aided thyroid nodule detection system using medical ultrasound images. Biomedical Signal Processing and Control, 40, pp.117-130.
20. Ma, J., Wu, F., Zhu, J., Xu, D. and Kong, D., 2017. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics, 73, pp.221-230.
21. Moradi, P. and Rostami, M., 2015. Integration of graph clustering with ant colony optimization for feature selection. Knowledge-Based Systems, 84, pp.144-161.
22. Ouyang, F.S., Guo, B.L., Ouyang, L.Z., Liu, Z.W., Lin, S.J., Meng, W., Huang, X.Y., Chen, H.X., Qiu-gen, H. and Yang, S.M., 2019. Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules. European Journal of Radiology, 113, pp.251-257.
23. Shankar, K., Lakshmanaprabu, S.K., Gupta, D., Maseleno, A. and de Albuquerque, V.H.C., 2018. Optimal feature-based multi-kernel SVM approach for thyroid disease classification. The Journal of Supercomputing, pp.1-16.
24. Sivagaminathan, R.K. and Ramakrishnan, S., 2007. A hybrid approach for feature subset selection using neural networks and ant colony optimization. Expert systems with applications, 33(1), pp.49- 60.
25. Sollini, M., Cozzi, L., Chiti, A. and Kirienko, M., 2018. Texture analysis and machine learning to characterize suspected thyroid nodules and differentiated thyroid cancer: Where do we stand?. European journal of radiology, 99, pp.1-8.
26. Tabakhi, S., Moradi, P. and Akhlaghian, F., 2014. An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence, 32, pp.112-123.
27. Tsantis, S., Dimitropoulos, N., Cavouras, D. and Nikiforidis, G., 2009. Morphological and wavelet features towards sonographic thyroid nodules evaluation. Computerized Medical Imaging and Graphics, 33(2), pp.91-99.
28. Wu, Y., Yue, X., Shen, W., Du, Y., Yuan, Y., Tao, X. and Tang, C.Y., 2013. Diagnostic value of diffusion-weighted MR imaging in thyroid disease: application in differentiating benign from malignant disease. BMC medical imaging, 13(1), p.23.
29. Yong, Z., Dun-wei, G. and Wan-qiu, Z., 2016. Feature selection of unreliable data using an improved multi-objective PSO algorithm. Neurocomputing, 171, pp.1281-1290.
30. Zhang, R., Huang, G.B., Sundararajan, N. and Saratchandran, P., 2007. Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 4(3), pp.485-495.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Advanced Machine Learning Techniques To Handle Brain Image Segmentation And Tumor Classification Over Bio-Medical Images

Mr .M AHARONU
Assistant Professor
Dept of CSE
Malla Reddy College Of Engineering
Hyderabad
Email:aharon.mattakoyya@gmail.com

Mrs. V Devi Priya
Assistant Professor
Dept of CSE
Malla Reddy College Of Engineering
Hyderabad
Email:vaddi.devipriya@gmail.com

Abstract

In real time applications, to evaluate mathematical closely related relations Rough set theory, Fuzzy set theory and rough set theory are the mathematical linear tools for uncertain data elements. Some of the researchers introduced rough sets, rough sets and fuzzy set by connected and combining all set theories together. In this research, we discuss about different combined notations of fuzzy, rough and rough set theories, and also discuss basic methods used to describe about above set theories effectively. We present the concepts related to rough based Rough intuitionistic fuzzy sets, intuitionistic fuzzy rough sets and discuss about basic properties of those set theories effectively. Furthermore, we discuss about classical presentation of rough based intuitionistic fuzzy sets in detail with approximate operations in real time synthetic applications. Segmentation of magnetic resonance images is medically complex and important for study and diagnosis of medical brain images, because of its sensitivity in terms of noise for brain medical images. These are the main issues in classification of brain images. Because of uncertainty & vagueness of brain medical images, so that rough sets, fuzzy sets and Rough sets are mathematical tools evaluate and handle uncertainty and vagueness in medical brain images. Traditionally, different type of fuzzy sets, Rough sets and rough sets based approaches were introduced, they have different several drawbacks with respect to different parameters. This research introduces a novel image segmentation (Classification) calculation method i.e. Enhanced and Explored Intuitionistic Rough based Fuzzy C-means Approach (EEISFCMA) with Support Vector machine classifier to estimation of weight bias parameter for brain image segmentation. Intuitionistic

Rough b ased fuzzy sets are generalized form of fuzzy, rough sets and their representative elements are evaluated with non-membership and membership value. Proposed algorithm of this paper consists standard features of existing clustering without spatial weight context data, it defines sensitive of noise in brain images, so that our proposed algorithm deals with intensity and noise reduction of brain image effectively. Furthermore, to reduce iterations in clustering, proposed algorithm initializes cluster centroid based on weight measure using max-dist evaluation method before execution of proposed algorithm. Experimental results of proposed approach carried out efficient image segmentation results compared to existing segmented approaches developed in brain image and other related images. Mainly proposed approach have consists better experimental evaluation based on results.

1. Introduction

In recent times, the introduction of information technology and e-health care system in the medical field helps clinical experts to provide better health care to the patient. Brain tumors affect the humans badly, because of the abnormal growth of cells within the brain. It can disrupt proper brain function and be life-threatening. Two types of brain tumors have been identified as benigntumors and malignanttumors. Benign tumors are less harmful than malignant tumors as malignant are fast developing and harmful while benign are slow growing and less harmful. The various types of medical imaging technologies based on noninvasive approach like; MRI, CT scan, Ultrasound, SPECT, PET and X-ray [1]. When compared to other medical imaging techniques, Magnetic Resonance Imaging (MRI) is majorly used and it provides greater contrast images of the brain and cancerous tissues. Therefore, brain tumor identification can be done through MRI images [2]. This paper focuses on the identification ofbrain tumor using image processing techniques. The detection of a brain tumor at an early stage is a key issue for providing improved treatment. Once a brain tumor is clinically suspected, radiological evaluation is required to determine its location, its size, and impact on the surrounding areas. On the basis of this information the best therapy, surgery, radiation, or chemotherapy, is decided. It is evident that the chances of survival of a tumor-infected patient can be increased significantly if the tumor is detected accurately in its early stage [3]. As a result,

the study of brain tumors using imaging modalities has gained importance in the radiology department. In this paper the brain tumor identification is done by an image processing. In this paper, there are four process are done to identify the brain tumors. The first process is pre processing the image data from the collection of database using median filtering, second stage is segmentation using Fuzzy C-means Clustering Algorithm [4], third stage is feature extraction using Gray Level Co-Occurrence Matrix (GLCM), [5] and the fourth stage is classification using ensemble classifiers is the combination of neural network, Extreme Learning Machine (ELM) and Support Vector Machine classifier (SVM). This will be discussed briefly in this following section.

2. Literature Survey

We can observe different data sets like Fuzzy sets, Rough sets and Soft sets notations with mathematical evaluations in real time applications based on different theories and developments. Present day's brain image segmentation is the basic problem to evaluate brain tumor decease in artificial intelligence real time applications. In medical image processing applications, brain tumor detection is a challenging task for real time medical applications. Traditionally some of the research authors introduced different machine learning methods, clustering approaches, classification approaches and filtering approaches to evaluate the basic procedure of the brain image segmentation in both theoretical and practical implementations based on above discussed data sets. All those approaches have some cons and pons in their implementations. In this section, we discuss about each technique implementations using real time data sets in image segmentation. Table 1 gives the brief discussion about all those techniques

Segmentation Approach	**Author**	**Description**	**Advantages**	**Disadvantages**
Adaptive Threshold	S. Jansi et.al	Based on image background, divide image into different dynamic regions based on threshold of various pixel	It will be worked based on thresholds	Less accuracy when rotation of different images applied, High time for processing images

		values		
K-Means Clustering	D.Selvaraj et.al	K-means clustering algorithm worked based on geometric interpretation of data. Based on centroid in images, it can identify brain tumor in images.	Less time for processing brain tumor segmentation, It is iterative process.	Less Accuracy, and Less false positive rate, not worked for large scale datasets
Improved K-Means Clustering	P. Vijayalakshmi et.al	Based on initial presentation of clustering identify brain tumor pixels in image segmentation.	It is easiest process, More accurate and high resolution	Less sensitivity and high time for image segmentation.
Fuzzy C-Means Clustering	M. Rakesh et.al	Based on given and pre-defined region and based on similarity measure identify brain tumor identification in images	More accurate in image segmentation	Give more time to identify tumor in brain images
Adaptive Fuzzy K-means Clustering	S. N. Sulaiman et.al	Based on degree measure relationship in images to identify brain tumor.	It is used to process Magnetic Resonance Images (MRI) Images	It is not applicable for qualitative and quantitative MRI brain images.
Region growing	Sudipta Roy et.al	Brain tumor identification is processed based on kindly segmentation	Extraction surface points may cardiac segmentation of	Requires user interface to formulate selection tumor presented surface from

ISBN 978-93-88808-61-3

		process applied on medication images	images	segmented images
Mean shift	Vishal B et.al	It is computer vision based non parametric clustering approach in medical image processing	It detect brain tumor on n-dimensional set presentation	Because of iterations in real time presentations, it computes high time complexity in segmentation
Watershed segmentation	Deorah et al	To identify foreground and background in image segmentation	Capturing of weak pixel formation in image segmentation, Less time for segmentation	Selection of Seed point selection is low, Increase convergence rate.
Level Set Model	Jiang Zhang et.al	To identify brain tumor in images based on surfaces at each dimension	Detection occurred based on level of surface identifications	It is not worked properly if curve was breaking.
K-Nearest Neighbour	Warfield et.al	Instance based brain tumor detection in brain image segmentation procedures	It is simplest approach to identify image segmentation, Increase accuracy	Statistical model to identify brain tumor presentations in brain images.
Support Vector Machine	Vapnik et.al	It is a supervised machine learning procedure to identify brain tumor presentation in image segmentation	It is an attractive and symmetric method to detect brain tumor image segmentation	Accuracy is very low in classification

Principal Component Analysis	Sumitra et al.	Based on principle feature presentation in images, identify the brain tumor classification in image segmentation	Reduce the large dimensionality in image segmentation	Less decomposition rate in image segmentations
Expectation maximization	Moon et al.	Based on some previously available tumor rules identify detection in brain image segmentation	Differentiate healthy and timorous tissues in brain image segmentation	It have intensity distribution of brain images.
Hierarchical clustering	Kshitij et al	Based on grouped tree clustering, identify tumors in brain images.	Accuracy is very high	Time complexity is very low
Back Propagation Algorithm	Rumelhard, D et.al	This method works properly in feed forward network ro identify tumor in brain images	Time complexity is less and easily verifiable	Less accuracy with feature extraction based on signal waves

Motivation

Consider the preliminaries present in table 1, we focus on development of advanced techniques to identify brain tumor in brain images based on segmentation/other properties. Our research mainly implement false positive rate, less time complexity and increase the accuracy in brain image segmentation to get better performance results of detection brain tumor in brain images.

3. Problem Statement

Fuzzy sets, soft sets and rough sets are the effective data processing frameworks for decision making relative to information

processing systems, information retrieval and other conclusive relations present in data, especially in some types of uncertain data events. So it is an efficient concept to process and effective dealing to evaluate uncertain data with different parameters. Consistently, number of researchers or authors has been introduced number of techniques in practical and theory oriented applications.

Define and discuss about different concepts related to fuzzy sets, rough sets and soft sets theories with their implementation in various fields with existing literature. To further implementation of this work is to develop soft rough with intuitionistic fuzzy sets to generalize properties of real time applications like image segmentation in brain oriented applications. We extend our research to support different mathematical evaluations in uncertain data processing in brain image segmentation with practical implementation.

4. Proposed Methodology

Main objective parameters for defined function i.e. K(X, Y,) to minimize standard representations for c-means for image segmentation in brain medical images. First we take derived parameters of defined function K(X, Y,) with respect to membership parameters x_{ij}, cluster centroid v_i and biased field $_k$ setting them into 0 and results of estimation matrixes of X(membership matrix), Y(centroid matrix) and (bias matrix). Based on these estimated results, we form our novel calculation and compute the classification of tissue and bias function field. Newly generated function of proposed approach is

$$K(X, Y,) = \sum_{i=1}^{c}\sum_{k=1}^{n} x_{ik}^{m} \| g_k - y_i \|^2 + \sum_{i=1}^{c} \sum_{k=1}^{n} x_{ik}^{m} (g_k - y_i)^2$$

Where $x = \frac{\sum_k g_k}{n}$

Estimation of Bias field

Taking the derivative of K(X, Y,) with respect to $_k$ and assign them into 0 then we have

$$\frac{\partial K(X, Y,)}{\partial_k} = \sum_{i=1}^{c} \sum_{k=1}^{n} x_{ik}^{m} (g_k - y_i)^2$$

Second summation of k^{th} term with respect to $_k$ then us have the following expression

$$\sum_{i=1}^{c} x_{ik}^{m} g_k - \sum_{i=1}^{c} x_{ik}^{m} \cdot {}_k - \sum_{i=1}^{c} x_{ik}^{m} v_i = 0$$

Differentiating the distance expression, then we obtain following expression

$$g_k \sum_{i=1}^{c} x_{ik}^{m} - {}_k \sum_{i=1}^{c} x_{ik}^{m} - \sum_{i=1}^{c} x_{ik}^{m} v_i = 0$$

$$x^m v$$

Distance based gradient function for bias field function is as follows:

$$\frac{1}{\sum_{i=1}^{c} x_{ik}^m} \sum_{i=1}^{c} x_{ik}^m y_k$$

If $\alpha \in (0,1)$ is the weight of membership function, then generated bias data is $\alpha = 0.007$ and increase this from 0.001 to $\alpha_2, \alpha_3, \ldots, \alpha_{10}$.

Updated Centroid of Cluster

Again taking the derivative of $K(X, Y, \alpha)$ with respect y_i and setting results is zero, then generated function is

$$\frac{\partial}{\partial x_{ik}} \sum_{k=1}^{n} x_{ik}^m (g_k - \alpha_k - \alpha_k y_i) = 0$$

Where $y_i^* = \frac{\sum_{k=1}^{n} x_{ik}^m (g_k - \alpha_k)}{\sum_{k=1}^{m} x_{ik}^m}$ after solving the above equation.

Intuitionistic Fuzzy based Image Representation

Intuitionistic fuzzy sets [IFS] representation of image for image segmentation. The presented image consists N*M size and the , where i is in between1to N*M, then image X to be represented in IFS as follows:

$$X = \{\langle a_i, \mu(a_i), \nu(a_i)\rangle \mid a_i \in A\}$$

with $\pi(a_i) = 1 - (\mu(a_i) - \nu(a_i))$, where $\mu(a_i)$ is membership function and $\nu(a_i)$ is non-member function and $\alpha(a_i)$ is the mean pixel value of image. After evaluating fuzzy image representation update each cluster based on different pixel values of image.

Evaluation of Membership

After minimize the above equation, with different constraints using Lagrange multiplier calculation

$$L(X, Y, \alpha) = \sum_{k=1}^{c} \sum_{i=1}^{n} x_{ik}^m \| g_k - \alpha_k - y_i \|^2 - \lambda_1 \left(\sum_{k=1}^{c} x_{ik}^m - 1 \right)$$

After taking derivative of $L(X, Y, \alpha)$ with respect to x_{ik} and set result into zero, then we have

$$\frac{\partial K(X, Y, \alpha)}{\partial x_{ik}} = \frac{\partial}{\partial x_{ik}} \sum_{m=1}^{2} \sum_{m=1}^{m} x_{ik}^m \| g_k - \alpha_k - y_i \| - \lambda_{x_{ik}} \left(\sum_{m=1}^{m} x_{ik}^m - x_{ik}^* \right) = 0$$

After solving the above equations based on different parameters for membership parameter sequences can be re=written as follows:

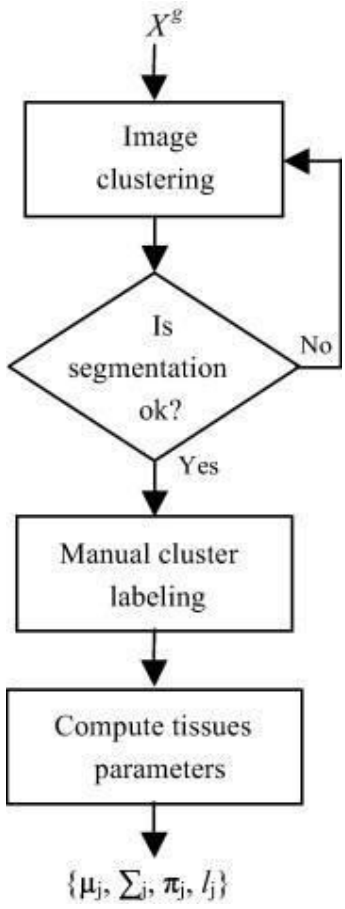
$$x_{ik}^* = \frac{1}{\sum_{j=1}^{c} \left(\frac{\| g_k - \alpha_k - y_i \|^2}{\| g_k - \alpha_k - y_i \|^2} \right)^{\frac{1}{m-1}}}$$

In the above equation c is number of

value of each pixel in image i.e.$A \square \{a_i / a_i\}$ centroid; g is gain function andis constant for membership function with different parameters.

EEISFCMA is evaluated on publicly available brain images, for example we collected brain images from https://www.nitrc.org/frs/?group_id=48&release_id=3124 and http://brainweb.bic.mni.mcgill.ca/brainweb/ with simulated brain image databases. We download these images from web urls and then convert into Matlab readable format and then we can pre-process for feature extraction to segment images using readable Rough ware i.e analysis and visualization of image. Proposed approach can be implemented in Latest Mat lab version with latest system configurations and this section describes implemented results. This section describes experimental results of different traditional approaches like k-means, fuzzy c-Mean, Generalized Fuzzy C-means, Gaussian Kernel based Fuzzy c-Means algorithm (GKFCM) and Rough fuzzy rough sets c-means (SFRCM) with proposed approach at segmentation accuracy and jacquard co-efficient for brain segmented images.

5. System Design



Design implementation of brain image segmentation for bio-medical images from different sources.

6. References

[1] Anupama Namburu, Srinivas kumar Samay, Srinivasa Reddy Edara, "Rough fuzzy rough set-based MR brain image segmentation",Proceedings in Applied Rough Computing xxx (2016) xxx–xxx. © 2016 Elsevier B.V. All rights reserved.

[2] S. Ramathilagam, R. Pandiyarajan, A. Sathya, R. Devi, S.R. Kannan," Modified fuzzy c-means algorithm for segmentation of T1–T2-weighted brain MRI", © 2010

Elsevier B.V. All rights reserved , Journal of Computational and Applied Mathematics 235 (2011) 1578–1586.

[3] K.V. Leemput, F. Maes, D. Vandermeulen, P. Suetens, Automated model based bias field correction of MR images of the brain, IEEE Trans. Med. Imaging 18 (1999) 885–896.

[4] M.N. Ahmed, N.A. Mohamed, A.A. Farag, T. Moriarty, A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data, IEEE Trans. Med. Imaging 21 (2002) 193–199.

[5] R.L. Cannon, J.V. Dave, J.C. Bezdek, Efficient implementation of the fuzzy c-means clustering algorithms, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8 (2) (1986) 248–255.

[6] Zujun Hou, A review on MR image intensity in homogeneity correction, Int. J. Biomed. Imaging (2006) 1–11. Article ID 49515.

[7] B.R. Condon, J. Patterson, D. Wyper, Image nonuniformity in magnetic resonance imaging: its magnitude and methods for its correction, Br. J. Radiol. 60 (1987) 83–87.

[8] M. Tincher, C.R. Meyer, R. Gupta, D.M. Williams, Polynomial modelling and reduction of spatial body-coil spatial in homogeneity, IEEE Trans. Med. Imaging 12 (1993) 361–365.

[9] S. Lai, M. Fang, A new variational shape-from orientation approach to correcting intensity inhomogeneities in MR images, in: Proc. of Workshop on Biomedical Image Analysis CVPR98, 1998, pp. 56–63.

[10] S.E. Moyher, D.B. Vigneron, S.J. Nelson, Surface coil MR imaging of the human brain with an analytic reception profile correction, J. Magn. Reson. Imaging 5 (1995) 139–144.

[11] S. Krinidis, V. Chatzis, A robust fuzzy localinformation c-means clusteringalgorithm, IEEE Trans. Image Process. 19 (5) (2010) 1328–1337.

[12] C. Li, J.C. Gore, C. Davatzikos, Multiplicative intrinsic component optimization(MICO) for MRI bias field estimation and tissue segmentation, Magn. Reson.Imaging 32 (7) (2014) 913–923.

[13] Z. Pawlak, Rough set approach to knowledge-based decision support, Eur. J.Oper. Res. 99 (1) (1997) 48–57.

[14] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17(2–3) (1990) 191–209.[15] P. Lingras, C. West, Interval set clustering of web users with rough k-means, J.Intell. Inf. Syst. 23 (1) (2004) 5–16.

[16] P. Maji, S.K. Pal, Rough set based generalized fuzzy c-means algorithm andquantitative indices, IEEE Trans. Syst. Man Cybern. B 37 (6) (2007)1529–1540.

[17] P. Maji, S.K. Pal, RFCM: a hybrid clustering algorithm using rough and fuzzysets, Fundam. Inform. 80 (4) (2007) 475–496

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

An Iot Based Approach For Energy Flexible Control Of Production Systems

P.Sandeep[1] Assistant Professor Malla Reddy College of Engineering
B.Shiva Karthik[2] Assistant Professor Malla Reddy College of Engineering

Abstract

Due to the increasing amount of renewable energy on the energy market resulting in a higher volatility of energy supply, manufacturing companies have an enhanced awareness of their energy demand in order to benefit from alternating prices. Energy flexibility is an opportunity to adapt manufacturing systems to the changing circumstances. The idea of energy flexibility follows the approach of synchronizing energy demand with supply, e.g. to exploit alternating weather conditions. This paper presents an energy-aware demand side management (DSM) approach to control manufacturing systems on the component level. The developed closed loop control is based on an algorithm fed with manufacturing, energy and environmental data and is realized at an Internet of Things (IoT) platform. Based on machine tool models the energy demand of a hypothetical factory is simulated. Taking on-site power generation data into account, the aim of the developed energy-aware control loop is to reduce the appearing residual power that must be balanced with grid-supplied power.

Keywords: energy flexibility; machine tools; on-site power generation; Internet of Things

1. Introduction

To achieve global climate agreements recently updated at the UN conference in 2015, new restrictions addressing the greenhouse gas (GHG) emissions were introduced by the German government. The Renewable Energy Law defines feed-in remuneration to increase the amount of renewable energy. As a result, the share of renewable energy has been increasing continuously to a rate of 29 % (188 GWh) in 2016 [1].

The German climate protection plan 2050 [2] includes a holistic energy concept addressing the energy sector, buildings, transport, agriculture and industry. For the industry sector, a reduction of GHG emissions of 49 % is striven for. Both the changing energy market with an increasing share of renewables and the rising viability of on-site power generation for manufacturing companies lead to a volatile energy supply. The adaption of the energy demand to supply plays a significant role to ensure competitiveness due to process stability, product quality and cost advantages.

This paper introduces an energy-conscious demand side management approach to control manufacturing systems on the component level. Based on a machine model the performance of the developed closed loop control is analyzed concentrating on the impact on CO_2 emissions, costs and time of grid neutrality. Furthermore, factory conditions for the application of the most suitable methods are identified and the approach is realized on an IoT platform.

2. Energy flexibility in smart factories

The future factory

Due to environmental circumstances, the entire factory structure will change resulting in new challenges. The conventional goal triangle in manufacturing companies is evolveing to a pyramid with the additional targets flexibility and sustainability [3] (figure 1).

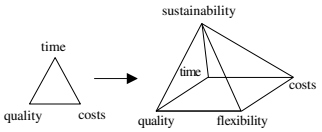


Fig. 1. Evolution of targets in manufacturing companies

Addressing sustainability, future factories should take social, economic and ecological aspects into account. A future concept of a sustainable manufacturing site is for example introduced by Stoldt et al. [4] with the key issues resource efficiency, zero emission and embedding people.

Besides the increasing awareness for sustainability, the digitalization influences the future factory significantly. The

advantages of the implementation of IoT technology in the future are to be found in literature:

- Flexibility, compatibility, scalability, ubiquity [5-8]
- Resource, cost and operational efficiency [7,8]
- Real-time capability and robustness [6,7,9]
- Usability and transparency [6,9]
- Complexity and intelligence [6,9]

The above-mentioned advantages of interlinking based on innovative information technology accelerates the fourth industrial revolution. Therefore, it is assumed, that the conventional automation pyramid will evolve to CPS (cyber-physical system) - based automation [10].

The implementation of IoT technologies supports the adaptation of future factories to changing environmental circumstances and can be useful for energy management in manufacturing companies.

Evaluation of energy flexibility in the research field of energy management in production systems

The research field of energy management includes different approaches and levels to face the challenges along with resource scarcity. Both energy data acquisition and analysis as well as energy data monitoring are requirements for energy flexible production planning or control. Overall energy management includes all aspects regarding resource allocation and planning. The evaluation of energy flexibility is observed on all re-search field levels.

In general, Reinhart et al. [11] define energy flexibility as *the capability of a production system to adapt quickly and with low financial expenditure to changes on the energy market*. Based on this definition, dimensions to identify the energy flexibility on the machine level are introduced [12]. Accordingly, energy flexible machines have low switching times, high power change rates and short critical times. Popp et al. [13] determine the degree of technical energy flexibility based on the components¶ demand and their relation among each other quantified with the Energy Independency Indicator (EII). Furthermore, energy flexibility indices are defined to evaluate energy flexibility on the component and on the machine level [14]. Simon et al. [15] introduce a method for the technical and economical evaluation of energy flexibility regarding the identification and categorization of measures.

The introduced evaluation approaches strive for energy flexible production planning and control. Beier et al. [16] present a detailed literature review of related research by dividing the relevant energy flexible research approaches into planning and real-time execution. Whereas the planning approaches include organizational methods, the real-time execution targets technical energy flexibility. Relevant technical research approaches are to be found in [13,16-23].

Data communication in energy flexible production systems

The implementation of IoT technology is in progress, thus different levels are covered in literature. The OPC UA interface commonly used in industry can be expanded for energy data transfer. Especially due to the platform-independency, the use of OPC UA is widespread [24]. Bauer et al. [25] abandon the hierarchical automatization pyramid. The concept targeting the adaption of energy demand to supply includes a so-called energy synchronization platform consisting of a market-side and a company-side platform. On the company-side platform the communication model is based on the paradigm *everything as a service*. A factory within this concept is already understood as a cyber-physical production system. Alternative approaches use wireless sensor networks to enable real-time energy monitoring [5,8]. Tan et al. [26] expand the approach of energy monitoring by a benchmark algorithm detecting advanced energetic statuses and conceptually introduce a totally IoT based approach. Shrouf et al. [27] develop an IoT energy management concept based on research, literature and expert interviews including both energy monitoring and a holistic integration of energy data into manufacturing.

3. IoT based closed loop control for energy flexible production systems

The introduced research works according to data communication in smart factories is currently on a conceptual level and not applied to energy flexible control approaches. Therefore, in the following an overall factory concept of an IoT based control loop is introduced and the simulation model structure, the control strategies and the model parametrization are defined and evaluated.

Overall concept of IoT based energy flexible factories

Figure 2 shows the overall factory model for the closed loop control for energy flexible production systems. Both on the component and on the factory level, demand data is measured and communicated to the cloud. Within the cloud, a database includes relevant energy information, e.g. the EII of all components. Furthermore, supply data from on-site generation and the power grid is provided to the cloud. To realize short-term prediction further data could be included, e.g. from weather or energy market forecasts. The implemented control strategy at the cloud computes the control commands from the given information according to the control strategy.

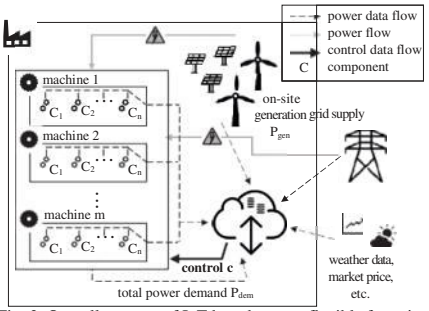


Fig. 2. Overall concept of IoT based energy flexible factories

Simulation model structure

To identify the impact of the closed loop control a simulation model was built up in Matlab Simulink. That model can

ISBN 978-93-88808-61-3

be executed locally or on the IoT platform ThingSpeak in an extended version. The modelling assumptions and simplifications were defined to detect relevant information only. The components¶ behavior is simulated with the following different modules (figure 3).

Functional storage module: Each component is modeled with a so-called functional storage, which is (un)loaded during the component¶V DFWLYH (passive) state. Based on the mean state time of the component, the storage size and the filling (emptying) gradient can be determined. The *internal* control switches the component to active (passive), when the storage reaches the bottom (top) dead center SOC_{bottom} (SOC_{top}).

Convergence module: This module balances the compo-QHQW¶V VWDte of charge (SOC)[1] at the end of the simulation to the start value (50 %) to avoid faults during the evaluation.

Reference component module: As a reference component module, a one-to-one copy of the introduced modules only with *internal* control was used to determine the differences between the only internal (storage-based) and the externally controlled (cloud-based) component.

To avoid inefficient control commands and high frequency switching, the external control is allowed in the following SOC range: $SOC \in \{10;90\}\ in\ \%$

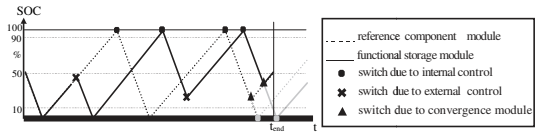


Fig. 3. 62& ¶LPH FRXUVH ZLWK VLPXODWLRQ PRGXOHV¶ IXQFWLRQV

In addition to the component subsystem, the model includes a determination subsystem, which computes the relevant key figures based on the input parameters

- x mean power demand in active state $p_{dem,a}$,
- x mean power demand in passive state $p_{dem,p}$,
- x component or machine status s
- x and the absolute SOC_{abs}.

The resulting key figures for the different control strategies are defined within formulas (1) to (3).

$$|\Delta p| \quad |p_{dem,a} - p_{dem,p}| \tag{1}$$

$$sign(\Delta p) \quad \begin{matrix} -,s\ active \\ +,s\ passive \end{matrix} \tag{2}$$

$$SOC \quad \frac{SOC}{SOC_{top} - SOC_{bottom}^{abs}} \tag{3}$$

To evaluate the impact of the developed energy flexible control strategies, present data were considered, whereas forecasts were neglected initially.

Control Strategies

To adapt the energy demand to the supply, three different control strategies are developed. All considered control strategies are based on the total power demand data (P_{dem}) and on-site generation data (P_{gen}). The difference between the two parameters is defined as the residual power (P_{res}), which is used to describe the interaction of the factory and the power grid (formula 4).

$$\begin{aligned} &P_{res} > 0,\ if\ P_{dem} > P_{gen} \text{ Æ } grid\ supply \\ &P_{res} < 0,\ if\ P_{dem} < P_{gen} \text{ Æ } grid\ feed\text{-}in \\ &P_{res} = 0,\ if\ P_{dem} = P_{gen} \text{ Æ } grid\ neutrality \end{aligned} \tag{4}$$

As third input, component data indices were used, whose specifications depend on the specific control strategy.

Strategy 1: power difference: The simplest decision rule is based on the FRPSRQHQW¶V PHDQ SRZHU GLIIHUHQFH $|\Delta p|$. The mean power differences of all regarded components are sorted by sign and by value. At first, all components with a mean power difference with the same sign as the residual power are excluded. Secondly, the largest remaining power difference is selected and the related component is switched (c = 1). Figure 4 shows the control strategy starting with the component with the maximum value of mean power difference.

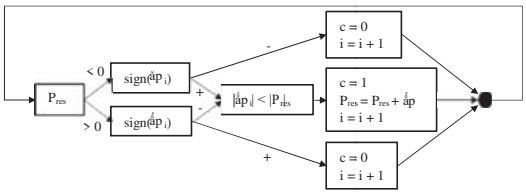


Fig. 4. Scheme of CS 1 (start: max ¸¨S_) DQG &6 2 (start: max/min SOC)

Strategy 2: state of charge: The SOC-control strategy follows the same scheme as strategy 1 (figure 5), but differs in iteration order. Whereas strategy 1 starts with the component i with the largest value of mean power difference, strategy 2 starts with the component holding the smallest/largest SOC.

Strategy 3: best fit: The third control strategy takes an additional static database into account, which includes all possible configurations of the system. For an exemplary five-component-system the corresponding database with all possible current states (rows) and all possible target states (columns) is computed resulting in a matrix with the dimension 2^5 x 2^5, since each component has two different states (active and passive). The matrix contains Δp between one current and one target state. Based on the value of the residual power the best fitting Δp is chosen to determine the target state. The method of control strategy 3 is shown in figure 5.

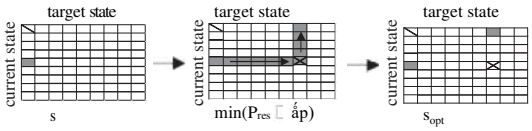


Fig. 5. Scheme of control strategy 3

Model implementation

The simulation model represents a virtual factory consisting of machines of three different types M_1 (4x), M_2 (2x)

[1] SOC = state of charge of the functional storage: term is used by extension to describe storages in general, e.g. electrical storages (battery back), thermal storages, pressure accumulators etc.

and M_3 (4x) and their energy independent components C_{11}, C_{12} (both M_1), C_2 (M_2) and C_3 (M_3). Measured power data of those components are provided in the component model. To consider all machine components, the total power demand on the factory level is based on measured data of five days and scaled regarding the installed amount of flexible energy. The on-site generation data is based on real measured data of radiation and wind during five days in November and scaled by the installed renewable power in the model.

4. Simulation procedure and method evaluation

The simulation model was used with different parametrization to analyze and evaluate the closed loop control considering three different purposes, explained in the following.

Simulation parametrization

Selecting the influencing parameters and configurations, the simulation model should lead to the identification of

- the performance of the control loop regarding the control strategies, the simulation step size and the delay time,
- the most suitable factory configuration considering the amount of energy flexible components and the dimensioning of installed on-site generation and
- the impact of the IoT environment.

Therefore, the simulation model ran according to the parameters shown in table 1.

Table 1. Simulation parameters

Parameter	Characteristics
database	average of a five-day-measurement of demand and on-site generation data
control strategy	CS 1, CS 2, CS 3, CS 12 (□å· CS 1, □å· CS 2), CS 23, CS 13, CS 123 ($^1/_3$ each)
step size	0.1 s, 0.5 s, 1 s
delay time	0.1 s, 15 s, 60 s
energy flexibility	10 %, 18 %, 25 %
Dimensioning of the on-site generation	1:0.5; 1:0.75; 1:1, 1:1.25, 1:1.25, 1:1.5, 1:2
model execution system	local, IoT

The introduced control strategies were applied individually (e.g. CS 1) or in combination by equal weight (e.g. CS 12). The *step size* is a simulation para-meter considering the size of simulation time steps and can be varied manually in the simulation. To ensure model plausibility the parameter specification for the *step size* was chosen in a certain range. The *delay time* describes the time lag within the system which in general occurs in closed control loops. The values for this parameter were considered regarding the minimal *delay time* (due to the model at least as high as the chosen step size) and expected delays within the IoT simulation (higher, not exact computable delay due to communication interfaces). The amount of flexible energy was initially set to 18 % (common value for machine tools [13,28]) and varied up-/downwards. The dimensioning of the on-site generation (DOG) was realized regarding the amount of energy demand, i.e. in case of 1:0.5 the generated amount of energy of five days is half of the energy demand over the same period.

Definition of key performance indicators (KPI)

To evaluate the closed loop control, three different key performance indicators were defined. The determination of all KPIs is based on the resulting residual power with and without application of the developed control loop.

KPI 1: reduction of CO_2 emissions: KPI 1 determines the impact of the control method regarding CO_2 emissions. Grid supply is weighted with the German CO_2 emission factor of 527 g/kWh (power trade balance) [29], whereas on-site generated power is assumed to be renewable and is therefore emission-free.

KPI 2: additional time of grid neutrality: This KPI evaluates the influence of the closed loop control on the time of grid neutrality, i.e. all simulation time steps with $P_{res} = 0$.

KPI 3: cost reduction: KPI 3 considers the economic evaluation concerning the running costs. Due to the newest development within the EEG legislation towards market-regulated feed-in remunerations and the decreasing production costs of renewable energy, the consumption of own-generated power will get more viable in the future. To weight on-site generation and grid supply power, future prices are used according to scenario B in [28] (table 2).

Table 2. Future scenario for energy price development

	Future Scenario	Unit
	Mean energy price	(grid
supply)	0.16	¼/N:K
Feed-in-rewards	0.06	¼/N:K
Own energy production costs	0.05	¼/N:K

Performance of the closed loop control

The performance was evaluated considering three parameters: *step size*, *delay time* and *control strategy*. To analyze and compare their influences, a sensitivity analysis was carried out. Figure 6 shows the sensitivity of the three KPIs for the *step size* (left) and the *delay time* (right).

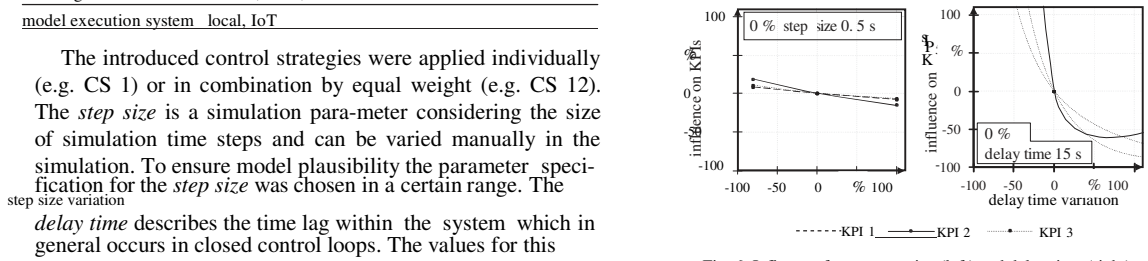


Fig. 6. Influence factors step size (left) and delay time (right)

Both parameters show an inversely proportional influence on the KPIs. Whereas the impact of a changing *step size* is very small, the variation of *delay time* shows a more distinct effect. The sensitivity for *step size* is approximately linear, i.e. in case of further increasing (decreasing) the *step size*, the effect on the KPIs gets equally smaller (higher). The highest sensitivity against the *step size* can be observed for KPI 2

(additional time of grid neutrality). In contrast, the observed impact declines very fast for increasing *delay time*. Nevertheless, a saturation is observed for KPI 2 and KPI 3, which means that delay times higher than a certain threshold do not further decrease the influence. The variation of the *delay time* changes the flexibility of the whole system, and therefore has a significant influence on all KPIs.

The results concerning the *control strategy* are shown in figure 7. The control strategies (x-axis) are sorted by their impact on the KPIs. Control strategy 2 shows the highest influence on all KPIs, whereas the lowest influence is observed for strategy 3. CS 1 is in the same range as CS 2. The insufficient results for control strategy 3 are explicable by the unconsidered input data SOC. In case of external control command in the EORFNHG FRPSRQHQW¶V 62& range, the computed best fit combination of the FRPSRQHQW¶V VWDWHs is not achieved. Both, CS 3 and CS 1, do not consider SOC as decision value. Nevertheless, the impact for strategy 3 is higher due to sequential formation of control commands. Executed simulations with combination CS23 and CS123 result in between the individual control strategies and are neglected in the presentation to ensure clear presentation.

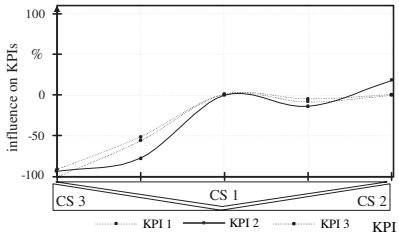


Fig. 7. Influence factor control strategy

To compare the parameters, the average of all KPIs was used to identify differences. Figure 8 shows the *delay time* with the largest impact on the KPIs.

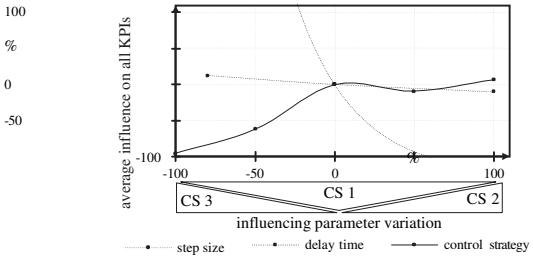


Fig. 8. Comparison of influence parameters

The *control strategy* cannot be treated as a continuous parameter. Comparing control strategy 1 and 2 the impact on the KPIs is as high as the influence of the *step size*. Strategy 3 shows an effect in the range of the *delay time*¶V LQI0XHQFH.

4.4. Factory configuration

The factory configuration was analyzed by regarding the parameters energy flexibility and dimensioning of on-site generation (DOG) as shown in figure 9. The impact of the energy flexibility is rising with increasing amount of flexible energy regarding KPI 1 and KPI 3. The influence of the energy flexibility on KPI 2 (additional time of grid neutrality) is very low in comparison. This can be explained by the variation of *energy flexibility* just based on the flexible energy and neglecting the flexible time of use, i.e. the period, the flexible energy is available. Therefore, increasing (decreasing) *energy flexibility* does not affect time parameters. Concerning the influence of *DOG*, the same effect was observable. Its impact on KPI 2 is lower than on KPI 1 or KPI 3, due to the dimensioning according to the energy amount only. The impact on the reduction of costs (KPI 3) shows a maximum at the dimensioning of 1:1.15. Based on the determination of the cost reduction a maximum close to a 1:1 was expected. The influence on KPI 1 (reduction of CO_2 emissions) is increasing with rising on-site generation. The differences in impact on KPI 1 and KPI 3 can be attributed to differences in their determination. Whereas KPI 1 (reduction of CO_2 emissions) weights grid-supplied power only, the determination of KPI 3 (reduction of costs) includes grid and self-supplied power.

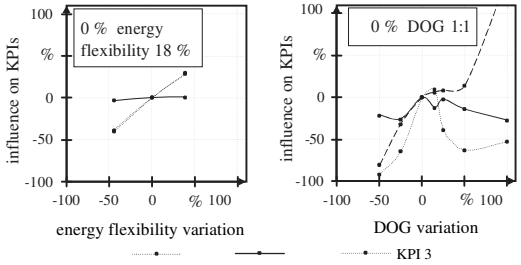


Fig. 9. Influence factors energy flexibility (left) and DOG (right)

4.5. IoT environment

To analyze the impact of the IoT environment, the local model was modified and partly implemented at the cloud. One of the main improvements of cloud-based closed loop control is the centralized accumulation of flexibility information, which is significant for decision making and developing control strategies for all different flexible components in a system of production machines to exploit all given energy flexibility potentials in an optimized way. The cloud-based simulation was executed with the following parameters:

control strategy 1, step size 0.1 s, delay time 0.1 s, energy flexibility 18 %, DOG 1:1. Figure 10 shows the result range of the locally conducted simulations and the IoT result range.

The results of the IoT simulation show conformity with the ORFDO VL**P**X0DWLRQ¶V UHVX0W UH**J**DUGLQ**J** KPI 1 and KPI 3. The IoT model outcome is approximately located in the middle for KPI 3, whereas the results for KPI 1 are in the lower edge. In case of KPI 2 the IoT simulation results do not reach the local simulation results. In IoT-based simulation the occurring delay times are higher than in locally execution and results in less sufficient performance regarding the KPIs. Nevertheless, the developed IoT model is applicable for the desired use case. Further analyses are in progress.

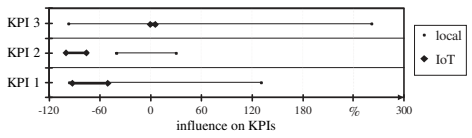


Fig. 10. IoT simulation results versus local simulation results

5. Conclusion and Outlook

The performance analysis indicates that the influence of the parameters differs. It is possible to deduce certain requirements for data communication in general. The delay time has a major influence on all considered KPIs. Therefore, it is important to provide data very fast, whereas the topicality of the data is less important. The results show that data conduction plays a significant role compared to data computing. The control strategies are able to reduce costs and CO_2 emissions and increase the time of grid neutrality. Nevertheless, control strategy 3 shows weak results compared to control strategy 1 or 2. The factory configuration has a higher input than the regarded influencing parameters. Therefore, it is important to implement energy flexibility and on-site generation in early planning steps and apply the closed loop control in addition to ensure most sufficient results. In addition to the conducted simulations further analyzes will be carried out with the IoT model to detect barriers and advantages of the cloud environment. Furthermore, the introduced IoT control loop will be integrated into machine tools to analyze the behavior under real conditions. An IoT communication system is already implemented and will be completed with the closed loop control for flexible production machines.

Acknowledgements

The authors would like to thank the German Federal Ministry of Education and Research (BMBF) and the Project Management Jülich (PtJ) for funding the project *SynErgie* (03SFK3E1).

References

[1] Umweltbundesamt. Erneuerbare Energien in Dtl. - Daten zur Entwicklung im Jahr 2016. Dessau-Roßlau/Germany, 2017. ISSN 2363-829X.

[2] Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit (BMUB). Klimaschutzplan 2050 ± Klimaschutzpolitische Grundsätze und Ziele der Bundesregierung. Berlin/Germany, 2016.

[3] Salonitis K, Ball P. Energy efficient manufacturing from machine tools to manufacturing systems. Procedia CIRP 7 (2013), pp. 634±639.

[4] Stoldt J, Franz E, Schlegel A, Putz M. Resource networks: Decentralised factory operation utilising renewable energy sources. Procedia CIRP 26 (2015), pp. 486±491.

[5] Mourtzis D, Vlachou E, Milas N, Dimitrakopoulos G. Energy consumption estimation for machining processes based on real-time shop floor monitoring via wireless sensor networks. Procedia CIRP 57 (2016), pp. 637±642.

[6] Tao F, Zuo Y, Xu LD, Lv L, Lin Z. Internet of Things and BOM-based Life Cycle assessment of energy-saving and emission-reduction of products. IEEE Transactions on Industrial Informatics 10 (2014) Vol. 2, pp. 1252±1261.

[7] Xu W, Yao B, Fang V, Xu W, Liu Q, Zhou Z. Service-oriented sustainable manufacturing: Framework and methodologies. International Conference on Innovative Design and Manufacturing. August 13rd ± 15th, 2014, Montreal, Quebec/Canada. pp. 305±310.

[8] Lu X, Wang S, Li W, Jiang P, Zhang C. Development of a WSN based real time energy monitoring platform for industrial applications. Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD). May 6th ± 8th, 2015, Calabria/Italy. pp. 337±342.

[9] Shrouf F, Ordieres J, Miragliotta G. Smart factories in Industry 4.0: A review of the concept of energy management approached in production based on the Internet of Things Paradigm. Proceedings of the 2014 IEEE IEEM. December 9th ± 12th, 2014, Selangor/Malaysia. pp. 697±700.

[10] Kowalewski S, Bettenhausen KD. Thesen und Handlungsfelder ± Cyber-physical Systems: Chancen und Nutzen aus Sicht der Automation. VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik, 2013.

[11] Reinhart G, Reinhardt S, Graßl M. Energieflexible Produktionssysteme: Einführung zur Bewertung der Energieeffizienz von Produktionssystemen. Werkstatttechnik online 102 Vol. 9 (2014),pp. 622±628.

[12] Zäh MF, Fischbach CWP, Kunkel F. Energieflexibilität in der Produktion identifizieren: Maßnahmen zur Nutzung und Beurteilungsgrößen. ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb 108 (2013) Vol. 9, pp. 639±642.

[13] Popp RSH, Zaeh MF. Determination of the technical energy flexibility of production systems. Advanced Materials Research 1018 (2014), pp. 365±372.

[14] Popp RSH, Liebl C, Zaeh MF. A multi-level procedure to evaluate the energy flexibility potential of production machines. Procedia CIRP (2017), accepted paper.

[15] Simon P, Schultz C, Keller F, Glasschröder J, Reinhart G. Energieflexibilität in Produktionssystemen. 14. Symposium Energieinnovation. February 10th ± 12th, 2016, Graz/Austria.

[16] Beier J. Simulation approach towards energy flexible manufacturing systems. Cham/Switzerland: Springer, 2017.ISBN 978-3-319-46639-2.

[17] Daryanian B, Bohn RE, Tabors RD. Optimal demand-side response to electricity spot prices for storage-type customers. IEEE Power Engineering Review 9 Vol. 8 (1989), pp. 897±903.

[18] Wang X, Ding H, Qiu M, Dong J. A low-carbon production scheduling system considering renewable energy. Proceedings of IEEE International Conference on Service Operations, Logistics, and Informatics (SOLI). July 10th ± 12th, 2011, Bejing/China. pp. 101±106.

[19] Li L, Sun Z, Yang H, Gu F. Simulation-based energy efficiency improvement for sustainable manufacturing systems. Proceedings of the ASME 2012 International Manufacturing Science and Engineering Conference (MSEC). June 4th ± 8th, 2012, Notre Dame/USA. pp. 1033±1039.

[20] Li L, Sun Z, Tang Z. Real time electricity demand response for sustainable manufacturing systems: challenges and a case Study. 8th IEEE International Conference on Automation Science and Engineering (CASE). August 20th ± 24th, 2012, Seoul/Korea. pp. 353±357.

[21] Zhou Z, Li L. Real time electricity demand response for sustainable manufacturing systems considering throughput bottleneck detection. IEEE International Conference on Automation Science and Engineering (CASE). August 17th ± 20th, 2013, Madison/USA. pp. 640±644.

[22] Sun Z, Li L. Potential capability estimation for real time electricity demand response of sustainable manufacturing systems using Markov decision process. Journal of Cleaner Production 65 (2014), pp. 184±193.

[23] Schultz C, Sellmaier P, Reinhart G. An approach for energy-oriented production control using energy flexibility. Procedia CIRP 29 (2015), pp. 197±202.

[24] Osterfeld H, Klimm B, Langer S, Schultz C, Reinhart G. Mit OPC UA zur energieorientierten Produktionssteuerung. productivity 20 (2015), pp. 49±52.

[25] Bauer D, Abele E, Ahrens R, Bauernhansl T, Fridgen G, Jarke M, Keller F, Keller R, Pullmann J, Reiners R, Reinhart G, Schel D, Schöpf M, Schraml P, Simon P. Flexible IT-platform to synchronize energy demands with volatile markets. Procedia CIRP 63 (2017), pp. 318±323.

[26] Tan YS, Ng YT, Low JSC. Internet-of-Things enabled real-time monitoring of energy efficiency on manufacturing shop floors. Procedia CIRP 61 (2017), pp. 376±381.

[27] Shrouf F, Miragliotta G. Energy management based on Internet of Things: Practices and framework for adoption in production management. Journal of Cleaner Production 100 (2015), pp. 235±246.

[28] Popp RSH, Liebl C, Zaeh MF. Evaluation of the energy flexible operation of machine tool components. Procedia CIRP 63 (2017), pp. 76±81.

[29] Umweltbundesamt. Entwicklung der spezifischen Kohlendioxid-Emissionen des deutschen Strommix in den Jahren 1990 ± 2016. Climate Change (15). Dessau-Rosslau/Germany, 2017

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

A survey of cyber security operations based on Machine learning & Deep learning

k.venkateswarlu
computer science and engineering
MallareddycollegeofEngineering
bagivenky@gmail.com

J.Avinash
computer science and engineering
MallareddycollegeofEngineering
avinashnayakjadhav@gmail.com

***Abstract*— In past decade machine learning (ML) and deep learning (DL), has generated irresistible research interest and attracted unprecedented public attention. With the increasing integration of the Internet and social life, there is change in how people learn and work, but it also exposes them to serious security threats. It is a challenging task to protect sensitive information, data, network and computers connected systems from the unauthorized cyberattacks. For this purpose, effective cyber security is required. Recent technologies such as machine learning and deep learning are integrated with cyberattacks to provide solution to this problem. The paper surveys machine learning and deep learning in cyber security also it discusses the challenges and opportunities of using ML / DL and provides suggestions for research directions.**

Keywords- Cyber security, Machine learning, Deep learning, Intrusion detection.

I. INTRODUCTION

Presently system connected by internet, such as the hardware, software & data can be protected from cyberattacks by means of cyber security. Cybersecurity is a set of technologies and processes designed to protect computers, networks, programs and data from attacks and unauthorized access, alteration, or destruction. As threats become more sophisticated the most recent technologies such as Machine learning (ML) and deep learning (DL) are used in the cybersecurity community to leverage security abilities. Nowadays, cyber security is a stimulating issue in the cyber space and it has been depending on computerization of different application domains such as finances, industry, medical, and many other important areas [11]. To identify various network attacks, particularly not previously seen attacks, is a key issue to be solved urgently [1].

This paper deals with previous work in machine learning (ML) and deep learning (DL) methods for cybersecurity applications and some applications of each method in cyber security operations are described. The ML and DL methods covered in this paper are applicable to detect cyber security threats such as hackers and predators, spyware, phishing and network intrusion detection in ML/DL. Thus, great prominence is placed on a thorough description of the ML/DL methods, and references to seminal works for each ML and DL method are provided [1]. And discuss the challenges and opportunities of using ML / DL for cybersecurity.

The rest of the survey is organized as follows:
Section II tells about cyber security, Section III is composed of Machine learning, Section IV contains survey on Deep learning and Section V dedicated to similarities and differences between Machine learning & Deep learning.

II. CYBER SECURITY

Protection of networks, computer connected devices, programs, and data from malicious attacks or unauthorized access using set of technologies is known as cyber security. Cyber security can be commonly referred as information technology security. Information can be sensitive information, or other types of data for which unauthorized access leads to disaster. In the process of synchronizing with new upcoming technologies, security trends and threat intelligence cyber security are at high risk. However, it is essential to protect information and data from cyberattacks, to maintain cyber security.

A. Challengesofcybersecurity

There are many challenges in the field of cyber security. One of the most challenging elements of cybersecurity is the changing nature of security threats. Traditionally protecting the biggest known threats and not protecting systems against less dangerous risks was approach against maintaining cyber security.

Key challenges of cyber security are:

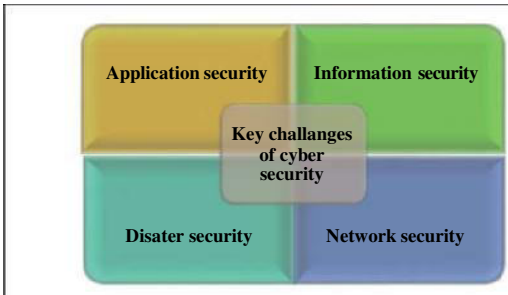


Fig.1. Areas covered in cyber security

- Application security: To protect applications from threats come from faults in the application design, development, deployment, upgrade or maintenance through actions that are taken during the development life-cycle is known as application security. Some basic methods used for application security are:
  1. Input parameter validation.
  2. User/Role Authentication & Authorization.
  3. Session management, parameter manipulation & exception management.
- Information security: It protects information from unauthorized access to save privacy. Methods used are:
  1. Identification, authentication & authorization of user.
  2. Cryptography.

- Disaster recovery planning: It is a process that comprises performing risk assessment, generating priorities, evolving recovery strategies in case of a disaster.
- Network security: Network security includes actions that are used to protect the usability, reliability, integrity and safety of the network. Security componentsinclude:
  1. Anti-virus and anti-spyware.
  2. Firewall, to block unauthorized access to your network.
  3. To identify fast-spreading threats, and Virtual Private Networks (VPNs) and to provide secure remote access intrusion prevention systems (IPS) is needed.

B. Typesofcybersecuritythreats

A cyberattack is a deliberate corruption of computers and servers, electronic systems, networks and data. Cyberattacks uses fake code to alter original computer code, logic or data, resulting in troublemaking consequences that lead to cybercrimes. End goal of cyber security is to prevent cyberattacks.

Following are some common types of cyber threats:

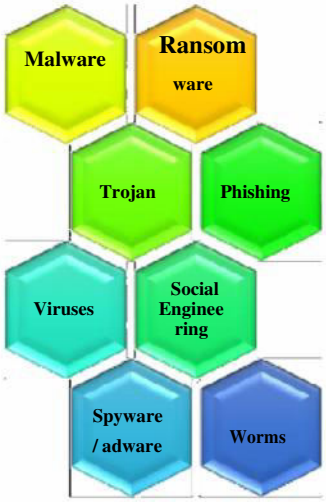


Fig.2. Types of cyber threats

- Type of activity that involves an attacker hacking system files through encryption and demanding a payment to decrypt is known as Ransomware.
- Malware is any file or program used to harm a computer user, such as worms, computer viruses, Trojan horses and spyware.
- Worms are like viruses in that they are self-replicating
- An attack that relies on human interaction to trick users for breaking security to gain sensitive is Social engineering.
- A virus is a piece of malicious code that is loaded onto a machine without the user's knowledge. It spread to other computers by attaching itself to another computerfile.
- Spyware/adware can be installed on computer without knowledge of user when attachments is opened or clicked or downloaded it infects the software and collects personal information.
- Trojan virus is performing malicious activity when executed.
- Phishing is a form of fraud where phishing attacks are sent via email and ask users to click on a link and enter their personal data. However, the intention of these emails is to steal sensitive data, such as credit card or login information. There is a concerning factor about phishing that phishing emails have become sophisticated and often look just like genuine requests for information.

III. MACHINE LEARNING

Machine learning (ML) allows software applications to predict outcomes without being explicitly programmed by use of an algorithm or group of algorithms. The machine learning builds algorithms for receiving input data and uses statistical analysis to predict an output while updating outputs as new data becomes available. Prior work in cyber security based on machine learning and artificial intelligence is presented below.

Liu et al., published a systematic study on security concerns with a variety of machine learning techniques. The existing security attacks explored towards machine learning from two aspects, the training phase and the testing/inferring phase [2]. Furthermore, categorization based on current defensive techniques of machine learning into security assessment mechanisms, countermeasures in the training phase, those in the testing or inferring phase, data security and privacy is done.

Paper presented by Fraley and Dr. Cannady gives better understanding of how machine learning could be leveraged to classify various security events and alerts. They developed model to react to security events by alerting SMEs, alerting analysts or producing reports depending upon the severity of the security event. Additional support for cyber defense was discussed to further reduce the time demand for responding to critical security events [3].

Merat et al. presented different types of computer processes that can be mapped in multitasking environment for the improvement of machine learning. SHOWAN model developed by them was used to learn the cyber awareness behavior of a computer process against multiple concurrent threads [4]. The examined process starts to outperform, and tended to manage numerous tasks poorly, but it gradually learned to acquire and control tasks, in the context of anomaly detection. Finally, SHOWAN plots the abnormal activities of manually projected task and compare with loading trends of other tasks within the group.

In the article, an overview of applying machine learning to address challenges in emerging vehicular networks was presented by Ye et al. This paper introduced basics of machine learning, including major categories and representative algorithms in brief. Some preliminary examples of applying machine learning in vehicular networks to ease data-driven decision making using reinforcement learning was published [5]. Some open issues for further research also highlighted in this paper.

A systematic of the challenges associated with machine learning in the context of big data and categorization based on the V dimensions of big data was published byL'Heureux

r [7]. An overview of ML approaches and how these techniques overcome the various challenges were discussed in this paper. The use of the big data to categorize the challenges of machine learning enables the creation of cause-effect connections for each of the issues. Further, the creation of explicit relations between approaches and challenges enables a more thorough understanding of ML with cyber security.

Golam et al., consider a data-driven next-generation wireless network model, where the MNOs employs advanced data analytics, ML and AI are used for efficient operation, control, and optimization. How ML, AI and computational intelligence play their important roles in data analytics for next-generation wireless networks are discussed in this paper. A set of network designs and optimization schemes with respect to data analytics are presented [8].

Feng and Wu presented a user-centric machine learning system which leverages big data of various security logs, alert information, and analyst insights to the identification of risky user. System provides a complete framework and solution to risky user detection for enterprise security operation center [12]. Generates labels from SOC investigation notes, to correlate IP, host, and users to generate user-centric features, to select machine learning algorithms and evaluate performances, as well as a machine learning system in SOC production environment was briefly introduced. The whole machine learning system is implemented in production environment and fully automated from data acquisition, daily model refreshing, to real time scoring, which greatly improve and enhance enterprise risk detection and management. As to the future work, learning algorithms was proposed for further improvement of the detection accuracy. Technological trends in anomaly detection and identification and open problems and challenges in anomaly detection systems and hybrid intrusion detection systems was discussed by Patcha et al. However, the survey only covers papers published from 2002 to 2006. Unlike Modi C et al., this review covers the application of ML / DL in various areas of intrusion detection and is not limited to cloud security. [1].
Buczak et al. proposed machine-learning methods and their applications to detect intrusion [1]. Algorithms like Neural Networks, Support Vector Machine, Genetic Algorithms, Fuzzy Logics, Bayesian Networks and Decision Tree are also described in paper.

Machine-learning methods are coarsely divided into three major categories as supervised, unsupervised, and reinforcement learning. There are two phases in machine learning i.e. training and testing. In the training stage, a model is learned based on training data, whereas in the testing stage, the trained model is applied to produce the prediction.

A. Supervised Learning

Supervised learning receives a labeled data set and further divide into classification and regression types. Each training sample comes with a discrete (classification) or continuous (regression) value called a label or ground truth. The goal of supervised learning is to gain the mapping from the input feature space to the label or decision space. Classification algorithms assign a categorical label to each incoming sample. Algorithms in this category include Bayesian classifiers, k-nearest neighbors, decision trees, support vector machines, and neural networks [5]. include logistic regression, support vector regression, and
the Gaussian process for regression [3].

B. Unsupervised Learning

For supervised learning, with enough data, the error rate can be reduced close to the minimum error rate bound. However, a large amount of labeled data is often hard to obtain in practice. Therefore, learning with unlabeled data, known as unsupervised learning, has attracted more attention. This method of learning aims to find efficient representation of the data samples, which might be explained by hidden structures or hidden variables, which can be represented and learned by Bayesian learning methods. Clustering is a representative problem of unsupervised learning, grouping samples into different clusters depending on their similarities. Input features could be either the absolute description of each sample or the relative similarities between samples. Classic clustering algorithms include k means, hierarchical clustering, spectrum clustering, and the Dirichlet process. Another important class of unsupervised learning is dimension reduction, which projects samples from a high-dimensional space onto a lower one without losing much information. In many scenarios, the raw data come with high dimension, and may want to reduce the input dimension for various reasons. In optimization, clustering, and classification, the model complexity and the number of required training samples dramatically grow with the feature dimension. Another reason is that the inputs of each dimension are usually correlated, and some dimensions may be corrupted with noise and interference, which will degrade the learning performance significantly if not handled properly

[5]. Some classic dimension reduction algorithms include linear projection methods, such as principal component analysis, and nonlinear projection methods, such as manifold learning, local linear embedding, and isometric mapping[5].

C. Reinforcement Learning

Reinforcement learning deciphers how to map situations to actions, through interacting with the environment in a trial-and-error search to maximize a reward, and it comes without explicit supervision. A Markov decision process (MDP) is generally assumed in reinforcement learning, which introduces actions and (delayed) rewards to the Markov process. The learning Q function is a classic model-free learning approach to solve the MDP problem, without the need for any information about the environment. This Q function estimates the expectation of sum reward when taking an action in a given state, and the optimal Q function is the maximum expected sum reward achievable by choosing actions. Reinforcement learning can be applied in vehicular networks to handle the temporal variation of wireless environments [5].

IV. DEEP LEARNING

Deep Learning is a sub area of Machine Learning research. It is a collection of algorithms in machine learning, used to model high-level abstractions in data. It Uses model architectures composed of multiple nonlinear transformations. Recently, it has made significant advances on various machine-learning tasks. Deep learning aims to understand the data representations, which can be built in supervised, unsupervised, and reinforcement learning. The input layer is at the leftmost, where each node in the figure

mension of input data. The output layer is at the rightmost, corresponding to the desired outputs, whereas the layers in the middle are called hidden layers. Typically, the number of hidden layers and the number of nodes in each layer are. A deep architecture means it has multiple hidden layers in the network as shown in figure 3. However, deeper networks bring new challenges, such as needing much more training data and gradients of networks easily exploding or vanishing. With the help of faster computation resources, new training methods (new activation functions, pretraining), and new structures (batch norm, residual networks), training such deep architecture becomes possible. Deep learning has been widely used in such areas as computer vision, speech recognition, and natural language processing and greatly improved state-of-the-art performance in these areas. Depending on applications, different structures can be added to the deep networks, e.g. convolutional networks share weights among spatial dimensions, whereas recurrent neural networks (RNNs) and long short-term memory (LSTM) share weights among the temporal dimensions [5].

Deep learning aims to learn a hierarchy of features from input data. It can automatically learn features at multiple levels, which makes the system be able to learn complex mapping function directly from data. The most characterizing feature of deep learning is that models have deep architectures. Deep architecture has multiple hidden layers in the network. In contrast a shallow architecture has only a few hidden layers (1 to 2 layers). Deep learning algorithms have been extensively studied in recent years. Algorithms are grouped into two categories based on their architectures:

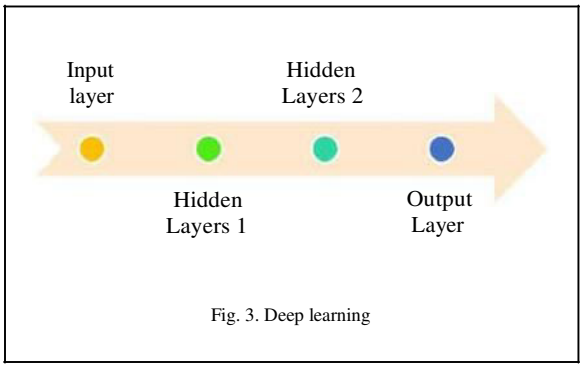


Fig. 3. Deep learning

A. Convolutionalneuralnetworks(CNN)

Convolutional neural networks (CNNs) has gain astonishing recognition in the field of computer vision. It has been continuously advancing the image classification accuracy. Also plays an important role for generic feature extraction such as scene classification, object detection, semantic segmentation, image retrieval, and image caption. Convolutional neural network (CNNs) is most important aspect of deep neural networks in image processing. It is highly effective and commonly used in computer vision applications. The convolution neural network composed of three types of layers: convolution layers, subsampling layers, and full connection layers.

B. RestrictedBoltzmannMachines(RBMs)

RBM is an energy-based probabilistic generative model. It is composed of one layer of visible units and one layer of hidden units. The visible units represent the input vector of a data sample and the hidden units represent features that are abstracted from the visible units. Each visible unit is connected to hidden unit, whereas no connection exists within the visible layer or hidden layer. During past years, the quality of image classification and object detection has been dramatically improved due to the deep learning method.

C. RecurrentneuralNetwork

RNNs are used to make use of sequential information. In a traditional neural network all inputs (and outputs) are independent of each other. To predict the next word in a sentence, need to know which words came before it. RNNs are called recurrent as they perform the same task for every element of a sequence, with the output being depended on the previous computations. RNNs can make use of information in arbitrarily long sequences, but in practice they are limited to only a few steps. An online unsupervised deep learning system is used to filter system log data for analyst. In which variants of Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs) are trained to recognize activity of each user on a network and concurrently assess whether user behavior is normal or anomalous, all in real time [10]. Developed model faced several key difficulties in applying machine learning to the cyber security domain. Model was trained continuously in an online fashion, but detection of malicious events was challenging task.

Comparative study was presented by Gavai et al. (2015) of a supervised approach and an unsupervised approach using the isolation forest method for detecting insider threat from network logs. Ryan et al. (1998) applied neural network-based approaches to train network with one hidden layer to predict the probabilities-based network intrusion [10]. A network intrusion was detected for the probability less than

But input features were not structured and did not train the network in an online fashion.

Modeling normal user activity on a network using RNNs was performed by Debar et al. (1992). The RNN was trained on a representative sequence of Unix command line arguments (from login to logout). Network intrusion detected when the trained network poorly predicts the login to logout sequence. While this work partially addresses online training, it does not continuously train the network to consider changing user habits over time.

Recurrent neural networks have been successfully applied to anomaly detection in various alternative domains such as signals from mechanical sensors for machinery such as engines, and vehicles [10].

An inclusive analysis of text Captchas, to evaluate security, a simple, effective and fast attack on text Captchas proposed by Tang et al. Using deep learning techniques, which successfully can attack all Roman character-based text Captchas deployed by the top 50 most popular websites in the world and achieved state-of-the-art results. Success rates range from 10.1% to 90.0% [9]. A novel image-based Captcha named SACaptcha using neural style transfer techniques also presented. This is a positive attempt to

security of Captchas by utilizing deep learning techniques. In this paper, deep learning techniques play two roles: as a character recognition engine to recognize individual characters and as a powerful means to enhance the security of the image-based Captcha This proved that deep learning is a double-edged sword. It can be either used to attack Captchas or improve the security of Captchas [9]. In future, they predicted existing text Captchas are no longer secure. Other Captcha alternatives are robust, and the designs of new Captchas can be simultaneously secure and usable are still challenging difficulties to be work on [9].

A new approach for detection of network intrusion using unsupervised deep learning with iterative K-means clustering proposed by Alom and Taha. In addition, unsupervised ELM, and only K-means clustering approaches were tested. From empirical evaluation on KDD-Cup 99 benchmark, it isobserved that the deep learning approach of RBM and AE with k-means clustering show around 92.12% and 91.86% accuracy for network intrusion detection respectively. RBM with K-means clustering provides around 4.4% and 2.95% better detection accuracy compare to K-means and USELM techniques respectively [11].

Nichols and Robinson present an online unsupervised deep learning approach to detect anomalous network activity from system logs in real time. Models decompose anomaly scores into the contributions of individual user behavior features for increased interpretability to aid analysts reviewing potential cases of insider threat. Using the CERT Insider Threat Dataset v6.2 and threat detection recall, their novel deep and recurrent neural network models outperform Principal Component Analysis, Support Vector Machine and Isolation [10].

V. SIMILARITIES AND DIFFERENCES BETWEEN MACHINE LEARNING & DEEP LEARNING

There are many puzzles about the relationship among ML, DL, and artificial intelligence (AI). Machine-learning is a branch of AI and is closely related to computational statistics, which also focuses on prediction making using computers [1]. whereas DL is a sub-field in machine-learning research. Its motivation lies in the establishment of a neural network that simulates the human brain for analytical learning. It mimics the human brain mechanism to interpret data such as images, sounds and texts [14].

A. Similarities

- Steps involved in ML and DL

ML and DL method primarily uses similar four steps in except feature extraction in DL is automated rather than manual [12].

- Methods used in ML and DL

ML/DL are similar in these three approaches: supervised, unsupervised and semi-supervised. In supervised learning, each instance consists of an input sample and a label. The supervised learning algorithm analyzes the training data and uses the results of the analysis to map new instances. Unsupervised learning that deduces the description of hidden structures from unlabeled data. Because the sample is unlabeled, the accuracy of the algorithm's output cannot be evaluated, and only the key features of the data can be summarized and explained. Semi-supervised learning is a means of combining supervised learning with unsupervised learning. Semi-supervised learning uses unlabeled data when using labeled data for pattern recognition. Using semi-supervised learning can reduce
label efforts while achieving high accuracy [1].

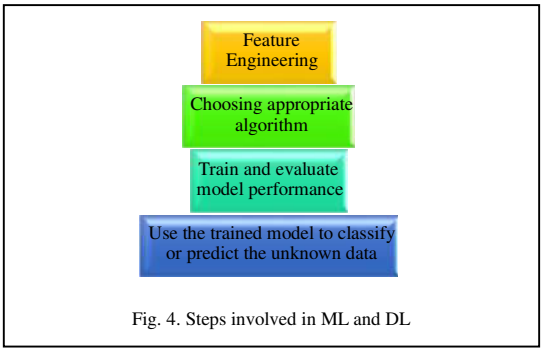


Fig. 4. Steps involved in ML and DL

B. Differences

ML and DL methods different in following ways:

- Data dependencies.

The main difference between deep learning and machine learning is its performance as the amount of data increases. Deep learning algorithms do not perform well when the data volumes are small, because deep learning algorithms require a large amount of data to understand the data perfectly. Conversely, machine-learning algorithm uses the established rules, thus performance is better.

- Hardware dependencies

The DL algorithm requires many matrix operations. The GPU is largely used to optimize matrix operations efficiently. Therefore, the GPU is the hardware necessary for the DL to work properly. DL relies more on high-performance machines with GPUs than machine-learning algorithms.

- Feature processing

The process of putting domain knowledge into a feature extractor to reduce the complexity of the data and generate patterns that make learning algorithms work better is known as feature processing. In ML, most of the characteristics of an application must be determined by an expert and then encoded as a data type. The performance of most ML algorithms depends upon the accuracy of the features extracted. Trying to obtain high-level features directly from data is a major difference between DL and traditional machine-learning algorithms. Thus, DL reduces the effort of designing a feature extractor for each problem.

- Problem-solving method

In Problem-solving method on applying traditional machine-learning algorithms to solve problems, traditional machine learning usually breaks down the problem into multiple sub-problems and solves the sub-problems, ultimately obtaining the result. Unlike deep learning which solves end-to-end problem.

- Execution time.

DL algorithm takes long time to train because there are many parameters in the DL algorithm. Whereas ML training takes relatively less time, only seconds to hours. The test time is exactly opposite for ML and DL. Deep learning algorithms require very little time to run during testing phase compared to ML algorithms. This is not applicable to all ML algorithms, some required short test times [1]

V. CONCLUSION

This paper provides researchers with a strong foundation for making easier and better informed choices about machine learning and deep learning for cyber security. It was reviewed that machine learning has some challenges in handling Big Data whereas deep learning performance is better in context of big data. To improve the security, an innovative image-based captcha named SACaptcha using deep learning techniques can be used. Unsupervised deep learning of RBM and AE with iterative k-means clustering show around 92.12% and 91.86% accuracy for network intrusion detection. In future, system of network intrusion detection for cyber security with online learning approach can be deployed. Machine learning is used to develop a model which detect and highlight advanced malware, by alerting SMEs, alerting analysts or producing reports depending upon the severity of the security event. The model performs these functions with very high accuracy (90%). To detect abnormal network activity from system logs in real time, an online unsupervised deep learning approach can be used that produces interpretable assessments of insider threat in streaming system user logs. This work has therefore accomplished its objective by providing with potential directions for future work and will hopefully serve as groundwork for great improvements of machine learning and deep learning methods for cyber security operations.

REFERENCES

[1] Yang Xin Et Al., "Machine Learning And Deep Learning Methods For Cybersecurity"In IEEE Journals & Magazine, May 2018.

[2] Qiang Liu Et Al., "A Survey On Security Threats And Defensive Techniques Of Machine Learning: A Data Driven View",In IEEE Journals & Magazine, Vol 6, February 2018.

[3] James B. Fraley And Dr. James Cannady, "The Promise Of Machine Learning In Cybersecurity", In Southeast conference , May 2017.

[4] Soorena Merat, P.Eng, Dr. Wahab Almuhtadi, P.Eng., "Artificial Intelligence Application For Improving Cyber-Security Acquirement" In 28th IEEE Canadian Conference On Electrical And Computer Engineering, Halifax, Canada, May 2015.

[5] Hao Ye, Le Liang et al., "Machine Learning for Vehicular Networks" In IEEE Vehicular Technology Magazine, April 2018.

[6] Ge Wang, Jong Chu Ye et al., "Image Reconstruction Is A New Frontier Of Machine Learning", In IEEE Transactions On Medical Imaging ,Vol. 37, pp 1289 – 1296, June 2018.

[7] Alexandra L'heureux et al., "Machine Learning With Big Data: Challenges And Approaches", In IEEE Journal & Magazine ,Vol 5, pp 7776 – 7797, April 2017.

[8] Mirza Golam Kibria Et Al., "Big Data Analytics, Machine Learning And Artificial Intelligence In Next-Generation Wireless Networks", In IEEE Journal & Magazine, May 2018, pp 2169-3536.

[9] Mengyun Tang Et Al., "Research On Deep Learning Techniques In Breaking Text-Based Captchas And Designing Image-Based Captcha", In IEEE Transactions On Information Forensics And Security,Vol13, Issue: 10,pp 2522 – 2537,Oct. 2018.

[10] Aaron Tuor ,Samuel Kaplan And Brian Hutchinson, "Deep Learning For Unsupervised Insider Threat Detection In Structured Cybersecurity Data Streams", In Proceedings Of Ai For Cyber Security Workshop At AAAI ,Dec 2017.

[11] Md Zahangir Alom And Tarek M. Taha, "Network Intrusion Detection For Cyber Security Using Unsupervised Deep Learning Approaches", In IEEE National Aerospace And Electronics Conference (NAECON), Dayton, Oh, USA, June 2017.

[12] Charles Feng*, Shunning Wu And Ningwei Liu, "A User-Centric Machine Learning Framework For Cyber Security Operations Center",
In IEEE International Conference On Intelligence And Security Informatics (ISI), Beijing, China, July 2017.

[13] Ozlem Yavanoglu And Murat Aydos, "A Datasets For Machine Learning Algorithms", In IEEE International
Conference On Big Data ,Jan 2018.

[14] A. Milenkoski, M. Vieira, S. Kounev, A. Avritzer, and B. D. Payne, "Evaluating Computer Intrusion Detection Systems:A Survey of Common Practices," AcmComput.Surv., vol. 48, no. 1, pp. 1–41, 2015.

[15] C. N. Modi and K. Acha, "Virtualization layer security challenges and intrusion detection/prevention systems in cloud computing: a comprehensive review," J. Supercomput., vol. 73, no. 3, pp. 1–43, 2016.

[16] E. Viegas, A. O. Santin, A. França, R. Jasinski, V. A. Pedroni, and L. S. Oliveira, "Towards an Energy-Efficient Anomaly-Based Intrusion Detection Engine for Embedded Systems," IEEE Trans. Comput., vol. 66, no. 1, pp. 163–177, 2017.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

An Environment Friendly System for Power Saving an Electrical Units by use of Image Processing

1. M.kranthi Assistant Professor Malla Reddy College of Engineering
2. K.Sudha Pavani Assistant Professor Malla Reddy College of Engineering

Abstract: The major problem in the most populated and developing countries like India, is Energy and Power crises. Hence there is a too much need of save energy. We use a several ways to save power like using the electric and electronic gadgets whenever and wherever needed and switching them off while not in use. But there are many places like classrooms, large auditoriums and meeting halls, there will be a fan or an Air-conditioner keeps running in unmanned area too, even before the people arrive. That improves the wastage of power in large amount and contributes to a considerable amount of electricity loss. As we all know about various methods for saving electricity like installing IR sensors to detect people etc. but it is quite costlier and complex in large areas. Here we propose a method of controlling power supply of auditoriums and classrooms using Image Processing. In this firstly we take reference image of empty classroom and if any changes in that reference image accordance with that power supply will turned on and off. This is very simple, efficient and cheaper technique to save energy. Also we include temperature sensor to sense temperature and calculate need of fan or equipments. Another big advantage is, we can extend this project up to application like home automation etc..

Keywords: Picture Processing, Image Partitioning, Edge Detection, Threshold Determination.

ISBN 978-93-88808-61-3

I INTRODUCTION

As we all know electricity is basic need of any business and we have to minimize any wastage of electricity. Video surveillance systems are widespread now a day. And it is widely use at airports, banks, casinos and correctional institutions. But now it is increased up to government agencies, business and even schools for increase security and provides video surveillance. As the availability of high speed, broad-band wireless networks and with the proliferation of inexpensive cameras, deploying a large number of cameras for security surveillance has become feasible with economical and technical manner. Several important research questions remain to be addressed before we can rely upon video surveillance as an effective tool for crime prevention, crime resolution and crime protection.

In video surveillance much of the current research focuses on algorithms to analyze video and other media from multiple sources for automatically detecting events. For eg. Intrusion detection, activity monitoring and pedestrian counting. Thus automated power management system is used to detect whether the room is empty or not. By using this technique we monitor changes in the auditorium through sequence of image and accordance with that the power supply is controlled. Image processing is a form of signal processing in that the input is an image and output may be either image or a set of characteristic related to the image.In this implementation first empty image is taken using digital camera then it is converted into gray and by using image enhancement technique we enhanced the image and apply edge detection. In the similar manner real time image is captured, enhanced and edge detected. Now the both images compared to each other and on the basis of results the control signal is generated by using hardware. Both images undergo the following processes

A. Acquisition
B. Gray Conversion
C. Partitioning
D. Edge Detection
E. Comparison
F. Generating control signal.

LITERATURE REVIEW

Literature survey is used to acquire knowledge and skill to complete this project. The main source for gaining the knowledge for this project is latest papers related on this topic. But there are some drawbacks of the previous research, to overcome that drawback and making the project more accurate we are doing several changes for making it more powerful algorithm.

By doing study on the previous research.the following conclusions are taken under consideration.

Accordingly to “Anisha Gupta/Punit Gupta2,jasmeet Chhabra” [3] they proposed intelligent automated system for an efficient power management is being deployed and tested over institutional building in which the lights of the classrooms are automatically controlled by the IOT device. That sense the real time occupancy based on the schedule uploaded on the database server,and takes intelligent action of controlling the lights of classrooms using electromagnet relay switch. The IOT device used here is Intel Galileo board and the sensor used for sensing the real time occupancy in motion detector sensor. The proposed system architectureis explained which include server connected to Intel Galileo board that automatically controls the lights of the class by realizing the real time occupancy of detecting the class using motion sensor[3].

With respect to “N. Sribhagat Verma, Ganesh Taduri ”[4] the need to automate the whole process of power management is very much there and this need is only going to escalate in the future with rising prices and scarcity of resources. Automated power management system is an effort in this direction and a small attempt to solve one of the biggest problems of mankind. With respect to our objective and scope,we have implemented and tested our system to the best possible. thus they calclude that automated power management system provides a practical and feasible approach to the problem of power management.[4]

“Kavya P. Walad, Jyoti Shetty ”[5] they discussed about existing traffic control system and their drawback, to overcome from those drawback can build a flexible traffic light control system based on traffic density. To find traffic density edge detection technique can be used. the edge detection is a well known technique in image processing from identifying an image object, image segmentation, image enhancement. Each edge detection technique have its own advantages and disadvantages in various fields. Gradients based or first order edge detection and Laplace based or second order edge detection operators are discussed in this paper can be implemented in MATLAB. There are so many drawbacks with Gaussian based edge detection is sensitive to noise. This is because of using static dimension of kernel filer and its coefficients. The canny edge detection gives the best performance even in noise condition compared to the first order edge detection. This is more costly compared to the Sobel. Prewitt and Robert’s operator. The main disadvantage with canny is that it has high computational time and responsible for weak edges. The best edge

detection technique is necessary to provide an errorless solution. In future rather than using existing edge detection technique can use fuzzy logic and morphological based edge detection technique for regulating traffic control system based on traffic density to save the time and reduce operating cost.

Accordingly "Manoj Kumar Asst. Professor, Dept of CSE"[6] they calculated all the various steps done and various results are compared with test cases. Students can be at corner or they can be at in front in a group etc. Test case I display two students are sitting and their subtracted image is another image also test case II display two students are sitting and their subtracted image is shown in another image. The study shows that this method is helpful in saving electricity. This method is very cheap, efficient and can reduce wastage of power. This will consistently detect that is there any person in a classroom and auditorium and hence saves electricity.[6] Accordingly "Vankatesh K and Sarath Kumar P "they conclude that image processing is better technique to control the power supply in the auditorium. It shows that it can reduce the wastage of electricity and avoids the free running of those electrical equipments. It is also more consistent in detecting presence of people because it uses real time images. Overall, the system is goog but it still needs improvement to achieve a hundred percent accuracy. If achieved, then we can extend this application to many places like theatres and even for home automation Also they proposed a scope for face detection.

With respect to "Shraddha Dhirde, Priyanka Ghuge, Sneha Khulape " they conclude that monitoring and controlling is done using parameter like temperature and human count by using Rasberry pi3. MB-LBP algorithm is implemented on the attributes of faces of people. This is one of the effective method to control the electric equipment and to reduce power consumption.

Kiiruthika G, Meenatchi R, Mohan raj[9] proposed a system that image processing is one of the useful technique to control the power supply in large areas like malls and auditorium. Also this prevents the free running of electrical application thereby reducing the power wastage. Also it proves to be a consistent and efficient technique to detect the presence of people since it uses real time image.

Patteri Sooraj, Faizankhan Pathan, Gohil Vishal[2] conclude that a classroom can be visualized where all the appliances can be controlled automatically without further human assistance. This makes the camera smart enough to monitor the electrical equipment and thus brings the whole idea of automation into classroom. Hence a lot of efforts and resources can be conserved which can be utilized for different purpose.

Vankatesh K developed a system in that image processing is main keyword to monitor the classroom and control power supply. The drawback with this system is that, it can be used only for the places whose orientation or arrangement is fixed. But they overcome it by resetting the reference images whenever the arrangement is altered. The main program needs not to be altered. Another way of overcoming this limitation is using face detection technique. That is expected to give much flexibility to the overall system.

For overcoming the previous problems related to the work, Here in this recent work we are using same technique of image processing with the temperature sensor and light sensor to sense the atmospheric temperature and light for calculating the need of appliances and making the system more accurate and convenient. In the alternation of face detection we are calculating the centred of object and on the basis of results, the operation will perform through microcontroller programming. In that we firstly take a reference image of empty classroom. This reference image compared to real time image after every 10 seconds. And with respect to changes, the operation will perform. There are many steps and parameter involves in this project that make it better and accurate than before.

CONCLUSION

The study showed that image processing is better technique to control the power supply in the classroom.This shows that it can minimize the wastage of electricity and avoid the free running of equipments. Also by using real time image we make it more consistent in detecting presence of people. Also by adding temperature and light sensor we make this system more accurate and convenient to use.

REFERENCES

[1] Vankatesh K1 and Sarath Kumar P2. "Automatic Real Time Auditorium Power Supply Control Using Image Processing". DOI:03.LSCS.2013.

[2] Patteri Sooraj1, Faizankhan Pthan2, Gohil Vishal3, Pritesh Kukadia4,"Automatic Controlling Of Electrical Appliances in Classroom Using Image Processing".IJERD:April 2016.

[3] Anisha Gupta/Punit Gupta2, Jasmeet Chhabra3."IoT based Power Efficient System Design Using Automation For Classroom". IEEE 2015.

[4] N. SribhagatVarma, Ganesh Taduri, N. BhagirathSai."Automatic Electrical Appliances control Based on Image Processing". IJERT September 2013.

[5] Kavya P Walad,JyothiShetty "Traffic Light Control System Using Image Processing". IJIRCCE October 2014.

[6] ManojKumar."Power Saving in Electrical Devices Using Image Processing". In IJEAST May 2016.

[7] OmkarRamdasGaikwad, Anil Vishwasrao, Prof. KanchanPujari, TejasTalathi "Image Processing Based Traffic Control"in IJSETR April2014.

[8] Shraddha Dhirde1, Priyanka Ghuge2, Sneha Khulape3 "Monitoring and Controlling of an Auditorium " in IJEST 2017.

[9] Kiruthika G1, Meenatchi R1,Mohan Raj V1, Pradeepa S1."Device Power Consumption avoidance using Image Processing" in IJARMET March 2017.

[10] Sunil Kumar Matangiand, SateeshPrathapani, "Design of Smart Power Controlling and Saving System in Auditorium by using MCS 51

[11] VikramadityDangi, AmolParab, KshitijPawar, S.S Rathod,"Image Processing Based Intelligent Traffic Controller", Undergraduate Academic Research Journal(UARJ), ISSN:2278-1129,volume-1, Issue-1,2012.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Sixth sense technology: Comparisons and future predictions

Sushmitha Chigiri*,Jhansi Rani Marla**
Department of Computer Science & Engineering,
MALLA REDDY COLLEGE OF ENGINEERING
(JNTU),
Hyderabad.

Abstract:
Sixth sense technology is a wearable gestural interface that enhances the physical world around us with the digital information and lets us use natural hand gestures to interact with that information.Sixth sense technology has integrated the real world objects with digital world. it associates technologies like hand gesture reorganization, image capturing, processing, and manipulations etc.

Key words:
sixth sense, hardware components, applications, advantages of sixth sense technology, future predictions in sixth sense technology.

Introduction:
The sixth sense technology is a mini projector coupled with a camera and a cell phone –which acts as the computer and connect to the cloud. All the information is stored on the web. Sixth sense is a new and interesting type of technology which is very easy to use by all the people. Sixth sense is a wearable gestural interface that augments physical word around us with the digital information. Sixth sense technology is a magic where each and everyone can use anywhere we want. It is a gateway between digital and real world.

Evolution:
Sixth Sense was developed at MIT Media Lab by Steve Mann in 1994 and 1997 (head worn gestural interface), and 1998 (neck worn version), and further developed by Pranav Mistry (also at MIT Media Lab), in 2009, both of whom developed both hardware and software for both head worn and neck worn versions of it. It comprises a head worn or neck-worn pendant that contains both a data projector and camera.

Sixth sense:
Every one of us are aware of five basic senses namely seeing, feeling, smelling, tasting, hearing. These senses have evolved millions of years ago .when these senses are not able to do any type of thing then sixth sense came it is just depend on our thinking. In this the information is stored on a paper or a digital storage device. Sixth sense device consists of projector, camera, mirror, web enabled phone and colored markers which are used to track our hand gestures by user hands.

Sixth sense device components

1. Camera
A camera is acting as a digital eye, which sees everything the user sees. The camera is meant to capture and recognize objects in its view and does the tracking of user's hand gestures using techniques based on computer-vision. The camera tracks all the movements made by the thumbs as well as the index fingers of both the hands of the user. On recognizing the object, the camera sends the data to a smart phone for processing.

2. Mobile Component
The sixth sense setup consists of an internet-enabled Smartphone which processes the data send from the camera. Smartphone is used to send and receive data and voice information from anywhere and to anyone through mobile internet. Software is run on the Smartphone which supports this technology and handles data connection. The Smartphone is meant to search the web and to interpret hand gestures. Computer–vision based techniques include programming using Symbian C++ code with more 50,000 lines of code.

3. Projector
The Smartphone interprets the data and this data is projected onto a surface mainly walls, body or hands of a person. A battery is found inside the projector which provides 3 hours battery life. Visual information is projected on to the surfaces and other physical objects which are used as interfaces by the projector. This projection of information is done by a tiny LED projector. The image is projected on to the mirror by the downward facing projector. On touching an object, the information related to the same will appear which will look like the information is part of the object.

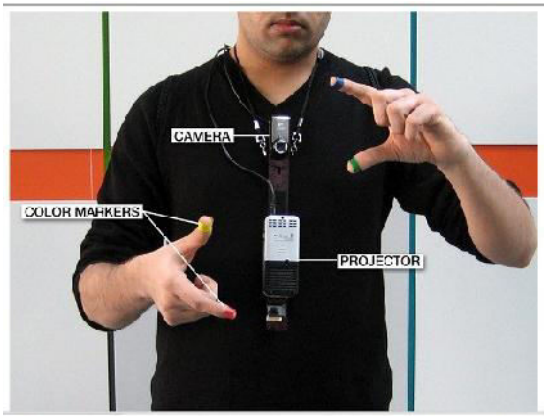
4. Mirror
The mirror is used as the projector hangs from the neck pointing downwards and it reflects the image to a desired surface. This step finally frees the

ISBN 978-93-88808-61-3

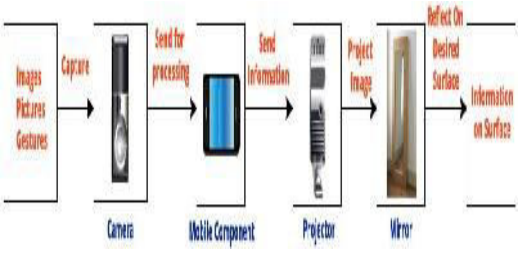
digital information from its confines and places it to the physical world.

5. Color Markers

The color markers that are red, green, blue and yellow are placed at the tips of the fingers which helps the camera to recognize the hand gestures. The various movements and structural arrangements made by these markers are interpreted as gestures that subsequently act as an instruction for the application interfaces are projected.



Working



- The sixth sense implementation hardware is a pendant like wearable mobile interface.
- It consists of a camera, Bluetooth or 3G or Wi-Fi enabled mobile component, projector, mirror and color markers.

Applications: The applications of sixth sense technology are so wide. As already stated, devices with this technology is meant to bring the digital information to the physical real world thereby bridging the existing gap. The recent sixth sense prototype device has showed off its usefulness, flexibility and viability of this technology. The only limitation to this technology is human imagination. Some of the practical applications of this technology are stated below:

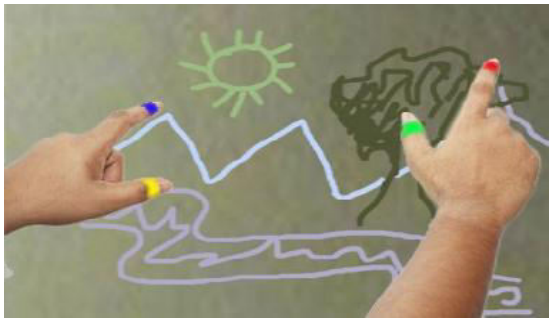
1. Taking Pictures

With the help of framing gestures by hands, user can take pictures of different locations in minimum time. After taking the photos, the user can review the pictures by displaying it onto any surface and then sorting, organizing and resizing the pictures.

2.Viewing Map

Navigation using maps are becoming very common these days. From millions of sources to millions of destinations, this application provides an accurate route along your path. At any instance, Map application helps the user to view any specific location and navigate through it by projecting the map onto a surface. With the help of fingers, mainly the thumb and index fingers, user can zoom in/out or pan the selected area.

3.Drawing Application



This application allows the user to draw on any surface and the drawings are tracked by the movements of finger tips especially the index finger. These pictures can be stored and replaced to any other surface with ease. User with the help of hand gestures can do shuffling through available pictures and drawings.

4.Making Calls

The sixth sense technology supporting device makes calling an easier job. This will project a keypad on your palm or use virtual keypad to make calls thus protecting the privacy. This technique is implemented in other technologies like Skin put. This application helps people with disability to call to a particular number at ease.

5.Interaction with physical objects

The sixth sense technology brings information about different physical objects in minimum time and in a better format. By drawing a circle on the wrist hand, displays an analog watch. Likewise, while reading newspapers, in place of written article, it shows live video news or else, even a paper is capable of providing dynamic information.

6.Grab Information

This technology driven devices are capable of providing information related to any object that is in touch with the user. For example, on holding a book, this device supplies the Amazon or Google ratings of the book as well as the reviews and other relevant details about the book. Also, using this technology, there is no more delay in searching the flight status. The device recognizes the boarding pass and informs the user about the flight status whether it is on time or not.

7.Sixth sense technology does some actions by simple customized hand gestures like an '@' simple when drawn will automatically redirect the user to check mails or a magnifying glass when drawn provides a map onto the surface.

Advantages

- A Sixth sense device has greatest advantage of having a small size and hence it is portable. All the components that make up this device are of light weight and the mobile component fits easily in user's pocket.
- Sixth sense technology comes with an added feature of multi-touch operation and multi-user interaction. Multi-touch operation allows multiple fingers of user to interact with device at a time.
- The prototype is a cost effective device. It consists of components that are common among other devices. The prototype costs around $350 and therefore it sure that when these devices come into the commercial

ISBN 978-93-88808-61-3

market on large scale the cost will become much lower.

- The device allows the user to access digital information at real time from any machine. Brain Computer Interface is not required to access the data.
- Sixth sense technology has broken the limits of a screen onto which a digital data can be projected and manipulated, that is now, data is available on any surface allowing a user to work on the same as per user's convenience.
- The devices are sure to change the habits of computer and machines and make it adaptable to that of humans since hand gestures captured by the device does jobs that was earlier done by machines.
- The software that supports this technology is likely to be an open source code as said by the developer.

Future Additions

As future enhancements, the team is working on to get rid of the use of color markers so as to capture gestures made by hands with ease. Also implementing camera and projector onto a Smartphone or mobile device together can reduce the total space occupied by both and will make it more handy. The team is also working on 3D gesture tracking since nowadays 3D images are a common scenario. This technology can be definitely a fifth sense for a disabled person.

Conclusion

Digital information nowadays is confined to the limits of computer screen or paper. Here, sixth sense technology is taking the natural ways of data display into a new phase which frees data from all its limits and integrates it with the real world seamlessly. It takes out the digital information to the physical world and bridges the gap by bringing the information from the intangible world to the tangible world. Sixth sense technology senses a physical object and projects all information about the object onto any surface let it be wall or hand/ body of a person. This technology is all set to bring that transparent interface to access information about anything and everything around us. Sixth sense technology is definitely the invention of the era and 'Get ready to be part of the magical world.

Reference:

http://www.pranavmistry.com
http://www.media.mit.edu
http://www.wikipedia.org
www.youtube.com
www.google.com
www.ted.com

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

A STATE OF THE ART OF DATABASES TO IMPROVE THE STORAGE EFFICIENCY ON CLOUD COMPUTING

Ch.Vengaiah[1] M.Ghanraj[2]

*T.Vijaykanth Reddy[3], SR.Mahipal[4]

Asst. Professor, Malla Reddy College of Engineering, Hyderabad

Asst. Professor, Malla Reddy College of Engineering, Hyderabad

*Research scholar, Saveetha School of Engineering, Saveetha University

*Research scholar, Hindustan instate of science and technology, Hindustan University

Abstract

The cloud computing is a large groups of remote servers are networked to allow the centralized data storage. It has the access of computer services, resources and can be classified as public, private and reserved. In this study, we explored various types of Data bases used in cloud computing with respect to the category of Knowledge database, XML database, Online databases and Real-Time databases to improve the storage and data efficacy.

Keywords: Cloud computing, Private Cloud, Public Cloud, Knowledge, Real-Time, Bibliographic Database, Bibliographic Database, mobile database.

1. INTRODUCTION

The database in cloud computing is categorized how it interact with various cloud sources for effectively improving the storage capacity for better performances. A data base is a organized collection of data are typically organized to model aspects of reality in a way that supports processes requiring information. For sample, modeling the availability of rooms in hotels in a way that supports finding a hotel with vacancies. And cloud computing is the computing in which large groups of remote servers are networked to allow the centralized data storing, and connected access to computer services or resources. Clouds can be categorizedby manner of public, reserved.

The data base in cloud computing is a storage architecture local administrator to cloud administrator. Traditional databases are organized by *fields*, *records*, and *files*. A field is a single piece of information; a record is one complete set of fields; and a file is a collection of records. To access information from a database, *database management system (DBMS)* is used. This is a collection of programs that supports you to enter, organize, and select data in a database.

The DB techniques are fundamental to increase data availability replication and synchronization. DB is divided into three levels such as front-end, middle-ware, and back-end that is built on Amazon Web Services front end and Mobiledevices is middle ware and also extensible markup language and back end cloud platform. It provides services and also high performance database process have seen exponential development in the past, and such growth is expected to quicken in the future.

2. RELATED WORK

A Cloud database management system (CDBMS) is a distributed database that delivers computing as a service instead of a product. It is the sharing of resources, software, and information between multiple devices over a network which is mostly the internet. Applications of database in real-time are Effective processing complex data

ISBN 978-93-88808-61-3

and data with set of the references for expression of the relations between them, Building of Internet-shops and distributed information systems, Building of the virtual company office and virtual kiosks, Storage and reproduction of graphic images, video and audio, Creation of WEB-sites, allotted to unlimited opportunities. Cloud applications connect to a database that is being run on the cloud and have varying degrees of efficiency. In this cloud computing the digital library books barrow problem is occur then they are kept application of homomorphic encryption mechanism for the library

The application of homomorphic encryption mechanism for the library of cloud computing Here they use collision data base[1]. The Design of an Adaptive Peer-to-Peer Network it reduce how it means the cloud of servers support thin clients with various types of service like Web pages and databases. Based on cloud computing peer to peer is now getting very popular

Such techniques are fundamental to increase data availability replication to synchronization have shown useful in the broad context of P2P systems and also super-peer collaborative systems Here they are using mobile data base [2].

Here the real time performs can be down that infrastructure only and Cloud-Mobile Computing Based Real-Time. Cloud-Mobile Computing Based Real-Time in this paper only we introduces a private cloud with SaaS service to realize a real-time video/voice over IP (VVoIP) [3] in this paper only we introduces a private cloud with SaaS they are having the huge capital investment.In their own IT infrastructure and also told that open environment where customers can deploy IT service providers may record service information in a service process from a customer and then collectively deduce the customer's private information

IT service providers may record service information in a service process from a customer and then collectively deduce the customer's private information [4].

Commonly here we are using with the help of computer and internet to get information based on cloud computing only we are share the resources only and also using Xml data base.And It is still in its infancy in regards to its Software as a Service (SaaS), Web Services, Utility Computing and Platform asService (PaaS).The location-based services and the abundant usage of smart phones and GPS-enabled devices. This is necessary to go that outsourcing data has grown rapidly over the past few years .cloud storage and cloud computing services has provided a flexible and cost-effective platform for hosting data from business and individuals in this knowledge-based development in the cloud rule engine and service oriented design in graph database has been designed to operate in a current cloud environment Cloud database are responsible for store data in high available form in cloud environment. The migration to one environment to another is difficult in that case cloud database uses to store and retrieve data

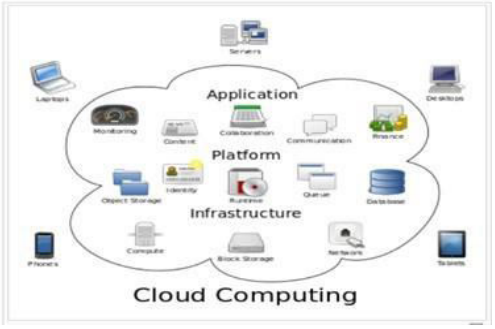


Figure 1. Cloud Computer metaphor: for user, network element

In fig1 cloud computer of metaphor user in network element is interacted with a applications, Infrastructure, and platform with different devices will be used. The privacy preserving system store data base of storage architecture local administrator to cloud administrator for this to learn about the outsourced database content and also more over the machine readable rights expressions are used in order to limit user of the database to a need-to-know basis Here they use cloud data base [5]

ISBN 978-93-88808-61-3

Cloud Storage for Real-Time Databases

Real-time Cloud Storage is a fast and fully managed backend-as-a-service (BaaS) that removes the administrative burden of operating distributed databases while providing seamless scalability. Designed for internet scale applications, Cloud Storage is particularly suited for online collaborative applications due to its powerful real-time notification features. Real time Cloud Storage is the ability of providing real-time notifications when data changes inside the storage. This means that's incredibly easy to develop applications that synch data between several users. Your application simply defines which events are of interest (e.g. table inserts, item updates, item deletes)

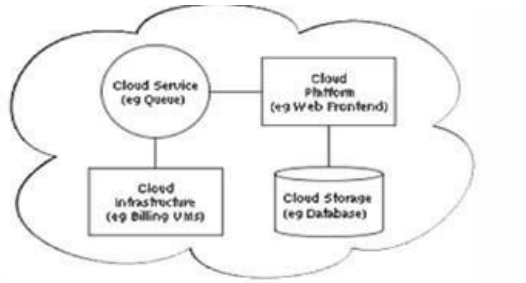


Figure 2.cloud computer sample architecture

Here the real time performs can be down that substructure only. In fig2 cloud computer example architecture A real time system can take advantage of intensive computing capabilities and scalable virtualized environment of cloud computing Here we are using Real-time data base [6] A real-time database is a database system which uses real-time processing to handle workloads whose state is constantly changing.

This differs from traditional databases containing determined data, mostly unaffected by time. For sample, a stock market changes very rapidly and is active. The graphs of the dissimilar marketplaces appear to be very unstable and yet a database has to keep track of current values for all of the markets of the New York Stock Exchange. Real-time processing means that a transaction is processed fast enough for the result to come back and be acted on correct away. Real-time databases are beneficial for accounting, multi-media, process control, and scientific data analysis

Cloud Storage for Knowledge Databases

Knowledge-based development approach for end-user in is a database used in the cloud environment. To practice the knowledge in the cloud rule engine and service oriented design were convoluted. It offers a framework for the user to store the knowledge, facts and actions. Here we use Knowledge database [7].

A Knowledge Database is a store of information that can be searched or browsed using pre-defined classifications. The classifications help to both guide the researcher and understand the context of the information they have found. Knowledge databases don't just leave users with a search box and let them work out what they should be looking for, knowledge databases provide knowledge database is a technology used to store complex structured and unstructured information used by a computer system.

The original use of the term knowledge-base was to describe one of the two sub-systems of a knowledge-based system. A knowledge-based system consists of a knowledge-base that represents facts about the world and an inference engine that can reason about those facts and use rules and other forms of logic to deduce new facts or highlight inconsistencies

- Flat data. Data was usually represented in a tabular format with strings or number in each field.
- Multiple users. A conventional database must support more than one user or system logged into the same data at the same time.
- Transactions. An essential requirement for a database was to maintain integrity and consistency among data that is accessed by concurrent users. These are the so-called ACID properties

Cloud Storage for XML Databases

Here using with the help of computer and internet to get information based on cloud computing only we are share the resources only and also using Xml data base .And It is still in its infancy in regards to its Software as a Service (SaaS), Web Services, Utility Computing and Platform as Service (PaaS) Here we use Xml database[8]. An **XML database** is a data persistence software system that allows data to be stored in XML format. These data can then be queried, exported and serialized into the desired format. XML databases are usually associated with document-oriented databases

- **XML-enabled**: these may either map XML to traditional database structures (such as a relational database[2]), accepting XML as input and rendering XML as output, or more recently support native XML types within the traditional database. This term implies that the database processes the XML itself (as opposed to relying on middleware).
- **Native XML** (NXD): the internal model of such databases depends on XML and uses XML documents as the fundamental unit of storage, which are, however, not necessarily stored in the form of text files.

XML in databases: the increasingly common use of XML for data transport, which has meant that "data is extracted from databases and put into XML documents and vice-versa".It may prove more efficient (in terms of conversion costs) and easier to store the data in XML format. In content-based applications, the ability of the native XML database also minimizes the need for extraction or entry of metadata to support searching and navigation. In a native XML environment, the entire content store becomes metadata through query languages such as X Path and XQuery, including content, attributes and relationships within the XML.

Cloud Storage for On-Line Databases In high performance database process databases have seen exponential growth in the past, and such growth is expected to accelerate in the future to increases the storage capacity comparing to old to implement like new thing Here we use the online database [9].

An online database is a database accessible from a network, including from the Internet.

It differs from a local database, held in an individual computer or its attached storage, such as a CD.

- For the system or software designed to Currently, there are several database products designed specifically as hosted databases delivered as software as a service products. These differ from typical traditional databases such as Oracle, Microsoft SQL Server, Sybase, etc. Some of the differences are:
- These online databases are delivered primarily via a web browser
- They are often purchased by a monthly subscription
- They embed common collaboration features such as sharing, email notifications, etc.

Cloud Storage for Bibliographic Database

In cloud computing research and selection system they are using the out ranking method because to get a better refine the results and also main contribution is conceiving an Agent that uses both the Skyline. Here we use Bibliographic databases[10].

The database of bibliographic records, an organized digital collection of references to published literature, including journal and newspaper articles, conference proceedings, reports, government and legal publications, patents, books, etc. In contrast to library catalogue entries, a large proportion of the bibliographic records in bibliographic databases describe articles, conference papers, etc., rather than complete monographs, and they generally contain very rich subject descriptions in the form of keywords, subject classification terms, or abstracts.

A bibliographic database may be general in scope or cover a specific academic discipline. A significant number of bibliographic databases are still proprietary, available by licensing agreement from vendors, or directly from the indexing and abstracting services that create them. Many bibliographic databases evolve into digital libraries, providing the full-text of the indexed contents. Others converge with non-bibliographic scholarly databases to create more complete disciplinary search engine systems, such as Chemical Abstracts.

Cloud Storage for mobile database

Here Designing and developing we use the three levels front-end, middle-ware, and a back-end that is built on Amazon Web Services front end is Mobil devise and middle ware is extensible markup language and back end cloud platform provides services Here we use theRelational Database [11].

A **mobile database** is either a stationary database that can be connected to by a mobile computing device (e.g., smartphones and PDAs) over a mobile network, or a database which is actually stored by the mobile device. This could be a list of contacts, price information, distance travelled, or any other information.[1]

Many applications require the ability to download information from an information repository and operate on this information even when out of range or disconnected. An example of this is your contacts and calendar on the phone. In this scenario, a user would require access to update information from files in the home directories on a server or customer records from a database. This type of access and work load generated by such users is different from the traditional workloads seen in client–server systems

Cloud Storage for collision Database

The Design of an Adaptive Peer-to-Peer Network it reduce how it means the cloud of servers support thin clients with various types of service like Web pages and databases. On based on cloud computing peer to peer is now getting very popular. Here we use collision database [12]. Collision induced absorption and emission refers to spectral features generated by inelastic collisions of molecules in a gas. Such inelastic collisions (along with the absorption or emission of photons) may induce quantum transitions in the molecules, or the molecules may form transient supra molecular complexes with spectral features different from the underlying molecules. Collision-induced absorption and emission is particularly important in dense gases, such as hydrogen and helium clouds in found in astronomical systems.

cloud storage for Time-series data base

In this real-time services they are having the huge capital investment in their own IT infrastructure and also told that open environment where customers can deploy IT service

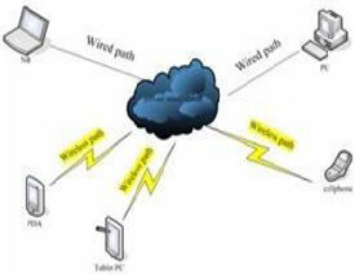


Figure 3.Cloud-mobile Computer via wired or wireless

providers may record service information in fig3 Cloud-mobile Computer via wired or wireless service process from a customer and then collectively deduce the customer's private information Here we use Time-series data base.

A time series database (TSDB) is a software system that is optimized for handling time series data, arrays of numbers indexed by time (a date time or a date time range). In some fields these *time series* are called profiles, curves, or traces. A time series of stock prices might be called a price curve. A time series of energy

ISBN 978-93-88808-61-3

consumption might be called a load profile. A log of temperature values over time might be called a temperature trace.

Despite the disparate names, many of the same mathematical operations, queries, or database transactions are useful for analysing all of them. The implementation of a database that can correctly, reliably, and efficiently implement these operations must be specialized for time-series data.

cloud storage for Spatial database

The location-based services and the abundant usage of smart phones and GPS-enabled devices. This is necessary to go that outsourcing data has grown rapidly over the past few years .cloud storage andcloud computing services has provided a flexible and cost-effective platform for hosting data from businesses and individuals Here we use Spatial database [13].

A spatial database, or geodatabase is a database that is optimized to store and query data that represents objects defined in a geometric space. Most spatial databases allow representing simple geometric objects such as points, lines and polygons. Some spatial databases handle more complex structures such as 3D objects, topological coverages, linear networks, and TINs. While typical databases are designed to manage various numeric and character types of data, additional functionality needs to be added for databases to process spatial data types efficiently. These are typically called geometry or feature. The Open Geospatial Consortium created the Simple Features specification and sets standards for adding spatial functionality to database systems

cloud storage for graph database

XGDBench is a graph database fig 4 Architecture of XGDB has been designed to operate in a current cloud environment. Cloud service benchmark to the domain of database bench-mark. It emphases on exascale cloud. This bench is centered on MAG model for realistic demonstrating of characteristic graphs Here we use graph database [14].

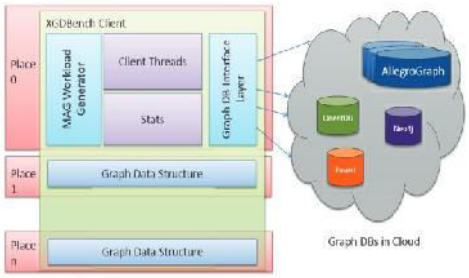


Figure 4. Architecture of XGDB

Graph databases have grown into increasingly popular for a variability of customs ranging from modeling to tracking software engineering enslavements in fig5 Virtual hierarchy as a graph. These extents use graphs because it expresses the awkward in graph traversal. Including migration this is used in hybrid cloud. It will provide a dramatic gain in concert. These databases solve the difficult in cloud management. The graph language database is very dominant. Here we use graph database [15]

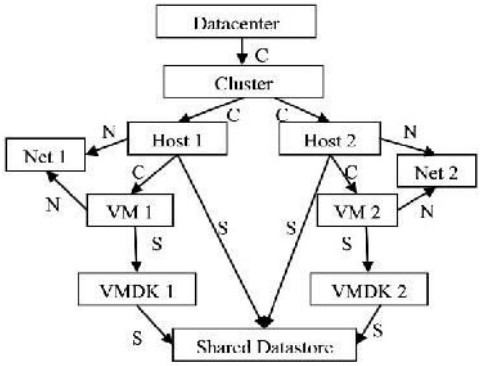


Figure 5. Virtual hierarchy as a graph

Cloud database are responsible for store data in high available form in cloud environment in fig5. The migration to one environment to another is difficult in that case cloud database uses to store and retrieve data. This provides a official way for drifting data among HBase as a column family database to Neo4j as graph databaseHere we use graph database [16].

3. OBSERVATIONS

1. Here we observe that the privacy preserving system store data base of storage architecture local administrator to cloud administrator
2. In this real time database we observe that intensive computer capabilities.
3. A private cloud with SaaS service to a real-time video, voice over IP.
4. To storing data in cloud computing is to get a better refine the result.
5. P2p system is a super-peer collaborative system.
6. In mobile device we use the three layers front end, middleware, backend for Designing and developing
7. The online database process is high performance have seen in exponential growth is past
8. In online database growth will be increases the storage capacity comparing to old.
9. Using the web pages of database we reduce servers support in thin client of peer to peer.

4. CONCLUSION

Finally we conclude the survey of database in cloud computing to improve the storage and data effectively. And here we use Various types of Data base in cloud computing like Bibliographic database, Knowledge database, XML database,Online databases, Real- time databases, Bibliographic Database.

5.REFERENCES

[1] Qingjie MENG, ChangqingGONG,Research of cloud computing security in digital library 6th International Conference on Information Management, Innovation Management and Industrial Engineering 2013.

[2] Barcelona, Spain,Data Replication and Synchronization in P2P Collaborative Systems 26th IEEE International Conference on Advanced Information Networking and Applications 2012.

[3] BaoRong Chang, Hsiu Fen Tsai, Chi-Ming, ChenYi-Sheng Chang, Chien-Feng Huang Cloud-Mobile Computing Based Real- Time VVoIP with PSO-ANFIS Tuning Conference on Technologies and Applications of Artificial Intelligence 2013.

[4] Gaofeng Zhang, Yun Yang, Xiao Liu, JinjunChen,A Time-series Pattern based Noise Generation Strategy for Privacy Protection in Cloud Computing 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing 2012. Computing,11th IEEE International Conference on Computer and Information Technology 2011.

[5] Rui Zhou, Jing Li, Jinghan Wang, GuoweiWang,A Knowledge-Based Development Approach with Fact and Service for End-User in Cloud Computing IEEE 37th Annual Computer Software and Applications Conference Workshops 2013.

[6] TarunKarnwal, T. Sivakuma, G. Aghila A Comber Approach to Protect Cloud Computing against XML DDoS and HTTP DDoS attack IEEE Students' Conference on Electrical, Electronics and Computer Science 2012.

[7] David Taniar,High Performance Database Processing 26th IEEE International Conference on Advanced Information Networking and Applications 2012.

[8] ManarABOUREZQ andAbdellah IDRISSI,Introduction of an outranking method in the cloud computing research and selection system based on the skyline.

[9] MaziyarShariatPanahi, Peter Woods, Peter Wood,Designing and developing a location-based mobile tourism application by using cloud-based platform International Conference on Technology, Informatics, Management, Engineering & Environment

(TIME-E 2013) Bandung, Indonesia, June 23-26,2013 2013.

[10] Apirajitha.P.S., AnithaAngayarkannA Design of an Adaptive Peer-to-Peer Network to Reduce Power Consumption Using Cloud Computing IEEE International Conference on Advanced Communication Control and ComputinTechnologies (ICACCCT) 2012.

[11] Ling Hu, Wei-Shinn Ku, SpiridonBakiras, Cyrus Shahab,Spatial Query Integrity with Voronoi Neighbors IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, no 4, April 2103.

[12] MiyuruDayarathna, ToyotaroSuzumura,XGDBench:

[13] VijayaraghavanSoundararajan and ShishirKakaraddi ,Applying Graph Databases to Cloud Management: An Exploration IEEE International Conference on Cloud Engineering 2014

[14] Mahdi NegahiShirazi, Ho Chin Kuan, HosseinDolatabadi,Design Patterns to Enable Data Portability between Clouds' Databases 12th International Conference on Computational Science and Its Applications 2012.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Securing Cloud Data under Key Exposure

Jangam Ravali,

M.Tech. Scholar,

Cse Department,
Malla Reddy College of Engineering

Abstract—Recent news reveal a powerful attacker which breaks data confidentiality by acquiring cryptographic keys, by means of coercion or backdoors in cryptographic software. Once the encryption key is exposed, the only viable measure to preserve data confidentiality is to limit the attacker's access to the ciphertext. This may be achieved, for example, by spreading ciphertext blocks across servers in multiple administrative domains—thus assuming that the adversary cannot compromise all of them. Nevertheless, if data is encrypted with existing schemes, an adversary equipped with the encryption key, can still compromise a single server and decrypt the ciphertext blocks stored therein. In this paper, we study data confidentiality against an adversary which knows the encryption key and has access to a large fraction of the ciphertext blocks. To this end, we propose Bastion, a novel and efficient scheme that guarantees data confidentiality even if the encryption key is leaked
and the adversary has access to almost all ciphertext blocks. We analyze the security of Bastion, and we evaluate its performance by means of a prototype implementation. We also discuss practical insights with respect to the integration of Bastion in commercial dispersed storage systems. Our evaluation results suggest that Bastion is well-suited for integration in existing systems since it incurs less than 5% overhead compared to existing semantically secure encryption modes.

Index Terms—Key exposure, data confidentiality, dispersed storage.

1 Introduction

THE world recently witnessed a massive surveillance program aimed at breaking users' privacy. Perpetrators were not hindered by the various security measures deployed within the targeted services [31]. For instance, although these services relied on encryp- tion mechanisms to guarantee data confidentiality, the necessary keying material was acquired by means of backdoors, bribe, or coercion.

If the encryption key is exposed, the only viable means to guarantee confidentiality is to limit the ad- versary's access to the ciphertext, e.g., by spreading it across multiple administrative domains, in the hope that the adversary cannot compromise all of them. However, even if the data is encrypted and dispersed across different administrative domains, an adversary equipped with the appropriate keying material can compromise a server in one domain and decrypt cipher- text blocks stored therein.

In this paper, we study data confidentiality against an adversary which knows the encryption key and has access to a large fraction of the ciphertext blocks. The adversary can acquire the key either by exploiting flaws or backdoors in the key-generation software [31], or by compromising the devices that store the keys (e.g., at the user-side or in the cloud). As far as we are aware, this adversary invalidates the security of most cryptographic solutions, including those that protect encryption keys by means of secret-sharing (since these keys can be leaked as soon as they are generated).

To counter such an adversary, we propose Bastion, a novel and efficient scheme which ensures that plaintext data cannot be recovered as long as the adversary has access to at most all but *two* ciphertext blocks, even when the encryption key is exposed. Bastion achieves this by combining the use of standard en- cryption functions with an efficient linear transform. In this sense, Bastion shares similarities with the no- tion of all-or-nothing transform. An AONT is not an encryption by itself, but can be used as a pre-processing step before encrypting the data with a block cipher. This encryption paradigm—called AON encryption— was mainly intended to slow down brute-force attacks on the encryption key. However, AON encryption can also preserve data confidentiality in case the encryp- tion key is exposed, as long as the adversary has ac- cess to at most all but one ciphertext blocks. Existing AON encryption schemes, however, require *at least* two rounds of block cipher encryptions on the data: one pre- processing round to create the AONT, followed by an- other round for the actual encryption. Notice that these rounds are sequential, and cannot be parallelized. This results in considerable—often unacceptable—overhead to encrypt and decrypt large files. On the other hand, Bastion requires only one round of encryption—which makes it well-suited to be integrated in existing dis- persed storage systems.

We evaluate the performance of Bastion in compari- son with a number of existing encryption schemes. Our results show that Bastion only incurs a negligible per-

- *G. Karame is affiliated with NEC Laboratories Europe, Heidelberg, 69115 Germany. E-mail: ghassan.karame@neclab.eu*
- *C. Soriente and S. Capkun are affiliated with the Compute Science Department of ETH Zurich, 8092, Switzerland. Email: first-name.lastname@inf.ethz.ch*
- *K. Lichota is affiliated with 9livesdata, Poland. Email: li- chota@9livesdata.com*

ISBN 978-93-88808-61-3

formance deterioration (less than 5%) when compared to symmetric encryption schemes, and considerably improves the performance of existing AON encryption schemes [12], [26]. We also discuss practical insights with respect to the possible integration of Bastion in commercial dispersed storage systems. Our contribu- tions in this paper can be summarized as follows:

- We propose Bastion, an efficient scheme which ensures data confidentiality against an adversary that knows the encryption key and has access to a large fraction of the ciphertext blocks.
- We analyze the security of Bastion, and we show that it prevents leakage of any plaintext block as long as the adversary has access to the encryption key and to all but two ciphertext blocks.
- We evaluate the performance of Bastion analytically and empirically in comparison to a number of existing encryption techniques. Our results show that Bastion considerably improves (by more than 50%) the performance of existing AON encryption schemes, and only incurs a negligible overhead when compared to existing semantically secure encryption modes (e.g., the CTR encryption mode).
- We discuss practical insights with respect to the deployment of Bastion within existing storage systems, such as the HYDRAstor grid storage system [13], [23].

The remainder of the paper is organized as follows. In Section 2, we define our notation and building blocks. In Section 4, we describe our model and introduce our scheme, Bastion. In Section 5, we analyze our scheme in comparison with a number of existing encryption primitives. In Section 6, we implement and evaluate the performance of Bastion in realistic settings; we also discuss practical insights with respect to the integration of Bastion within existing dispersed storage systems. In Section 7, we overview related work in the area, and we conclude the paper in Section 8.

2 PRELIMINARIES

We adapt the notation of [12] for our settings. We define a block cipher as a map $F : \{0, 1\}^k \times \{0, 1\}^l \rightarrow \{0, 1\}^l$, for positive k and l. If P_l is the space of all $(2^l)!$ l- bits permutations, then for any $a \in \{0,1\}^k$, we have $F(a, \cdot) \in P_l$. We also write $F_a(x)$ to denote $F(a, x)$. We model F as an ideal block cipher, i.e., a block cipher picked atrandom from $BC(k, l)$, where $BC(k, l)$ is the space of all block ciphers with parameters k and l. For a given block cipher $F \in BC(k, l)$, we denote $F^{-1} \in BC(k, l)$ as $F^{-1}(a, y)$ or as $F_a^{-1}(y)$, for $a \in \{0, 1\}^k$.

Encryption modes

An encryption mode based on a block cipher F/F^{-1} is given by a triplet of algorithms $\mathcal{Q} = (\mathsf{K}, \mathsf{E}, \mathsf{D})$ where:

- K The key generation algorithm is a proba- bilistic algorithm which takes as input a security parameter k and outputs a key $a \in \{0, 1\}^k$ that specifies F_a and F_a^{-1}.
- E The encryption algorithm is a probabilistic algorithm which takes as input a message $x \in \{0, 1\}^*$, and uses F_a and F_a^{-1} as oracles to output ciphertext y.
- D The decryption algorithm is a deterministic algorithm which takes as input a ciphertext y, and uses F_a and F^{-1} as oracles to output plaintext $x \in \{0, 1\}^*$, or $\perp$ if y is invalid.

For correctness, we require that for any key $a \leftarrow \mathsf{K}(1^k)$, for any message $x \in \{0, 1\}^*$, and for any $y \leftarrow \mathsf{E}^{F_a, F_a^{-1}}(x)$, we have $x \leftarrow \mathsf{D}^{F_a, F_a^{-1}}(y)$.

Security is defined through the following chosen-plaintext attack (CPA) game adapted for block ciphers:

$\mathbf{Exp}^{ind}(A, b)$
$F \leftarrow BC(k, l)$
$a \leftarrow \mathsf{K}(1^k)$
$x_0\, x_1\, state \leftarrow A^{\mathsf{E}^{F_a, F_a^{-1}}}(find)$
$y_b \leftarrow \mathsf{E}^{F_a, F_a^{-1}}(x_b)$
$b' \leftarrow A(guess, y_b, state)$

In the *ind* experiment, the adversary has unrestricted oracle access to $\mathsf{E}^{F, F^{-1}}$ during the "find" stage. At this point, A outputs two messages of equal length x_0, x_1, and some *state* information that are passed as input when the adversary is initialized for the "guess" stage (e.g., *state* can contain the two messages x_0, x_1). During the "guess" stage, the adversary is given the ciphertext of one message out of x_0, x_1 and must guess which message was actually encrypted. The advantage of the adversary in the *ind* experiment is:

$$\mathrm{Adv}^{ind}_{\mathcal{Q}}(\mathrm{A}) = |\Pr[\mathrm{Exp}^{ind}_{\mathcal{Q}}(A, 0) = 1] - \Pr[\mathrm{Exp}^{ind}_{\mathcal{Q}}(A, 1) = 1]|$$

Definition 1. An encryption mode $\mathcal{Q} = (\mathsf{K}, \mathsf{E}, \mathsf{D})$ is *ind* secure if for any probabilistic polynomial time (p.p.t.) adversary A, we have $\mathrm{Adv}^{ind}_{\mathcal{Q}}(\mathrm{A}) \leq \varrho$, where ϱ is a negligible function in the security parameter.

REMARK 1. The *ind* experiment allows the adversary to see the entire (challenge) ciphertext. In a sce- nario where ciphertext blocks are dispersed across a number of storage servers, this means that the ind- adversary can compromise all storage servers and fetch the data stored therein.

REMARK 2. In the *ind* experiment (and in other experiments used in this paper), we adopt the Shannon Model of a block cipher that, in practice, instanti- ates an independent random permutation for every different key. This model has been used in previous

ISBN 978-93-88808-61-3

related work [3], [12], [17] to disregard the algebraic or cryptanalysis specific to block ciphers and treat them as a black-box transformation.

All or Nothing Transforms

An All or Nothing Transform (AONT) is an efficiently computable transform that maps sequences of input blocks to sequences of output blocks with the following properties: *(i)* given all output blocks, the transform can be efficiently inverted, and *(ii)* given all but one of the output blocks, it is infeasible to compute any of the original input blocks. The formal syntax of an AONT is given by a pair of p.p.t. algorithms Q = (E, D) where:

- E The encoding algorithm is a probabilistic algorithm which takes as input a message $x \in \{0,1\}^*$, and outputs a pseudo-ciphertext y.
- D The decoding algorithm is a deterministic algorithm which takes as input a pseudo- ciphertext y, and outputs either a message $x \in \{0,1\}^*$ or $\perp$ to indicate that the input pseudo-ciphertext is invalid.

For correctness, we require that for all $x \in \{0,1\}^*$, and for all $y \leftarrow \mathsf{E}(x)$, we have $x \leftarrow \mathsf{D}(y)$.

The literature comprises a number of security definitions for AONT (e.g., [8], [12], [26]). In this paper, we rely on the definition of [12] which uses the *aont* ex- periment below. This definition specifies a block length l such that the pseudo-ciphertext y can be written as $y = y[1] \ldots y[n]$, where $|y[i]| = l$ and $n \geq 1$.

$\mathbf{Exp}^{aont}(A, b)$
$x, state \leftarrow A(find)$
$y_0 \leftarrow \mathsf{E}(x)$
$y_1 \leftarrow \{0,1\}^{|y_0|}$
$b' \leftarrow A^{Y_b}(guess, state)$

On input j, the oracle Y_b returns $y_b[j]$ and accepts up to $(n-1)$ queries. The *aont* experiment models an adversary which must distinguish between the encod- ing of a message of its choice and a random string (of the same length), while the adversary is allowed access to all but one encoded blocks. The advantage of A in the *aont* experiment is given by:

$$\mathrm{Adv}^{aont}_{\mathcal{Q}}(A) = |\Pr[\mathrm{Exp}^{aont}(A, 0) = 1] - \Pr[\mathrm{Exp}^{aont}(A, 1) = 1]|$$

Definition 2. An All-or-Nothing Transform $\mathcal{Q} = (\mathsf{E}, \mathsf{D})$ is $aont$-secure if for any p.p.t. adversary A, we have $\mathrm{Adv}^{aont}_{\mathcal{Q}}(A) \leq \varrho$, where ϱ is a negligible function in the security parameter.

Known AONTs

Rivest [26] suggested the *package transform* which leverages a block cipher F/F^{-1} and maps m block strings to $n = m + 1$ block strings. The first $n-1$ output blocks are computed by XORing the i-th plaintext block with $F_K(i)$, where K is a random key. The n-th output block is computed XORing K with the encryption of each of the previous output blocks, using a key K_0 that is publicly known. That is, given $x[1] \ldots x[m]$, the package transform outputs $y[1] \ldots y[n]$, with $n = m + 1$, where:

$$y[i] = x[i] \oplus F_K(i), \; 1 \leq i \leq n-1,$$
$$y[n] = K \bigoplus_{i=1}^{n-1} F_{K_0}(y[i] \oplus i).$$

Desai [12] proposed a faster version where the block cipher round which uses K_0 is skipped and the last output block is set to $y[n] = K \bigoplus_{i}^{n-1} y[i]$. Both AONTs are secure according to Definition 2 [12].

Remark 3. Although most proposed AONTs are based on block ciphers [12], [26], an AONT is not an encryption scheme, because there is no secret-key information associated with the transform. Given all the output blocks of the AONT, the input can be recovered without knowledge of any secret.

3 System and Security Model

In this section, we start by detailing the system and security models that we consider in the paper. We then argue that existing security definitions do not capture well the assumption of key exposure, and propose a new security definition that captures this notion.

System Model

We consider a multi-cloud storage system which can leverage a number of commodity cloud providers (e.g., Amazon, Google) with the goal of distributing trust across different administrative domains. This "cloud of clouds" model is receiving increasing attention nowa- days [4], [6], [32] with cloud storage providers such as EMC, IBM, and Microsoft, offering products for multi- cloud systems [15], [16], [29].

In particular, we consider a system of s storage servers $S_1, \ldots, S_s$, and a collection of users. We assume that each server appropriately authenticates users. For simplicity and without loss of generality, we focus on the read/write storage abstraction of [21] which exports two operations:

- write(v) This routine splits v into s pieces $\{v_1, \ldots, v_s\}$ and sends (v_j) to server S_j, for $j \in [1 \ldots s]$.
- read() The read routine fetches the stored value v from the servers. For each $j \in [1 \ldots s]$, piece v_j is downloaded from server S_j and all

ISBN 978-93-88808-61-3

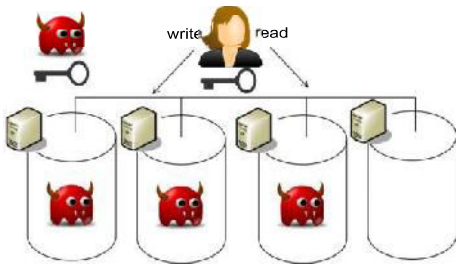


Fig. 1. Our attacker model. We assume an adversary which can acquire all the cryptographic secret material, and can compromise a large fraction (up to all but one) of the storage servers.

pieces are combined into v. We assume that the initial value of the storage is a special value $\perp$, which is not a valid input value for a write operation.

Adversarial Model

We assume a computationally-bounded adversary A which *can acquire the long-term cryptographic keys used to encrypt the data.* The adversary may do so either *(i)* by leveraging flaws or backdoors in the key-generation software [31], or *(ii)* by compromising the device that stores the keys (in the cloud or at the user). Since ciphertext blocks are distributed across servers hosted within different domains, we assume that the adversary cannot compromise all storage servers (cf. Figure 1).

In particular, we assume that the adversary can com- promise all but one of the servers and we model this adversary by giving it access to all but λ ciphertext blocks.

Note that if the adversary also learns the user's cre- dentials to log into the storage servers and downloads all the ciphertext blocks, then no cryptographic mech- anism can preserve data confidentiality. We stress that compromising the encryption key does not necessarily imply the compromise of the user's credentials. For example, encryption can occur on a specific-purpose device [10], and the key can be leaked, e.g., by the manufacturer; in this scenario, the user's credentials to access the cloud servers are clearly not compromised.

$(n-\lambda)$-CAKE Security

Existing security notions for encryption modes capture data confidentiality against an adversary which does not have the encryption key. That is, if the key is leaked, the confidentiality of data is broken.

In this paper we study an adversary that has access to the encryption key but does not have the entire ci- phertext. We therefore propose a new security definition that models our scenario.

As introduced above, we allow the adversary to access an encryption/decryption oracle and to "see" all but λ ciphertext blocks. Since confidentiality with $\lambda = 0$ isclearly not achievable[1], weinstead seekanencryption mode where $\lambda = 1$. However, having the flexibility of setting $\lambda \geq 1$ allows the design of more efficient schemes while keeping a high degree of security in practical deployments. (See Remark 7.)

We call our security notion $(n-\lambda)$ Ciphertext Access under Key Exposure, or $(n-\lambda)$*CAKE*. Similar to [12], $(n-\lambda)$*CAKE* specifies a block length l such that a ciphertext y can be written as $y = y[1] \ldots y[n]$ where $|y[i]| = l$ and $n > 1$.

$$\begin{aligned}&\mathbf{Exp}^{(n-\lambda)CAKE}(A, b)\\ &\quad a \leftarrow \mathsf{K}(1^k)\\ &\quad x_0\, x_1\, state \leftarrow A^{\mathsf{E}^{F_a, F_a^{-1}}}(find)\\ &\quad y_b \leftarrow \mathsf{E}^{F_a F_a^{-1}}(x_b)\\ &\quad b' \leftarrow A^{Y_b, \mathsf{E}^{F_a, F_a^{-1}}}(guess, state)\end{aligned}$$

The adversary has unrestricted access to $\mathsf{E}^{F_a, F_a^{-1}}$ in both the "find" and "guess" stages. On input j, the oracle Y_b returns $y_b[j]$ and accepts up to $n-\lambda$ queries. On the one hand, unrestricted oracle access to $\mathsf{E}^{F_a, F_a^{-1}}$ captures the adversary's knowledge of the secret key. On the other hand, the oracle Y_b models the fact that the adversary has access to all but λ ciphertext blocks. This is the case when, for example, each server stores λ ciphertext blocks and the adversary cannot compromise all servers. The advantage of the adversary is defined as:

$$Adv^{(n-\lambda)CAKE}_{\mathcal{Q}}(A) = Pr[\mathbf{Exp}^{(n-\lambda)CAKE}_{\mathcal{Q}}(A, 1) = 1] - Pr[\mathbf{Exp}^{(n-\lambda)CAKE}_{\mathcal{Q}}(A, 0) = 1]$$

Definition 3. An encryption mode = (K, E , D) is $(n-\lambda)$*CAKE* secure if for any p.p.t. adversary A, we have $\mathrm{Adv}^{(n-\lambda)CAKE}_{\mathcal{Q}}(A) \leq \varrho$, where ϱ is a negligible function in the security parameter.

Definition 3 resembles Definition 2 but has two fundamental differences. First, $(n-\lambda)$*CAKE* refers to a keyed scheme and gives the adversary unrestricted access to the encryption/decryption oracles. Second, $(n-\lambda)$*CAKE* relaxes the notion of all-or-nothing and parameterizes the number of ciphertext blocks that are not given to the adversary. As we will show in Sec- tion 4.2, this relaxation allows us to design encryption modes that are considerably more efficient than existing modes which offer a comparable level of security.

We stress that $(n-\lambda)$*CAKE* does not consider confi- dentiality against "traditional" adversaries (i.e., adver- saries which do not know the encryption key). Indeed, an *ind*-adversary is not given the encryption key but has access to all ciphertext blocks. That is, the *ind*- adversary can compromise all the s storage servers. An $(n-\lambda)$*CAKE*-adversary is given the encryption key but can access all but λ ciphertext blocks. In practice,

1. Any party with access to all the ciphertext blocks and the encryption key can recover the plaintext.

ISBN 978-93-88808-61-3

the $(n - \lambda)$*CAKE*-adversary has the encryption key but can compromise up to $s - 1$ storage servers. Therefore, we seek an encryption mode with the following properties:

1) must be *ind* secure against an adversary which does not know the encryption key but has access to all ciphertext blocks (cf. Definition 1), by compro- mising all storage servers.
2) must be $(n - \lambda)$*CAKE* secure against an ad- versary which knows the encryption key but has access to $n - \lambda$ ciphertext blocks (cf. Definition 3), since it cannot compromise all storage servers.

REMARK 4. Property 2 ensures data confidentiality against the attacker model outlined in Section 3.2. Nevertheless, we must also account for weaker ad- versaries (i.e., traditional adversaries) that do not know the encryption key but can access the entire ciphertext —hence, *ind* security. Note that if the adversary which has access to the encryption key, can also access all the ciphertext blocks, then no cryptographic mechanism can preserve data confi- dentiality.

4 BASTION: SECURITY AGAINST KEY EXPOSURE

In this section, we present our scheme, dubbed Bastion, which ensures that plaintext data cannot be recovered as long as the adversary has access to all but *two* ciphertext blocks—even when the encryption key is exposed. We then analyze the security of Bastion with respect to Definition 1 and Definition 3.

Overview

Bastion departs from existing AON encryption schemes. Current schemes require a pre-processing round of block cipher encryption for the AONT, fol- lowed by another round of block cipher encryption (cf. Figure 2 (a)). Differently, Bastion first encrypts the data with one round of block cipher encryption, and then applies an efficient linear post-processing to the ciphertext (cf. Figure 2 (b)). By doing so, Bastion relaxes the notion of all-or-nothing encryption at the benefit of increased performance (see Figure 2).

More specifically, the first round of Bastion consists of CTR mode encryption with a randomly chosen key K, i.e., $y' = Enc(K, x)$. The output ciphertext y' is then fed to a linear transform which is inspired by the scheme of [28]. Namely, our transform basically computes $y = y' \cdot A$ where A is a square matrix such that: *(i)* all diagonal elements are set to 0, and *(ii)* the remaining off-diagonal elements are set to 1. As we shown later, such a matrix is invertible and has the nice property that $A^{-1} = A$. Moreover, $y = y' \cdot A$ ensures that each input block y'_j will depend on all output blocks y_i except from y_j. This transformation—combined with the fact that the original input blocks have high entropy (due to semantic secure encryption)—result in an *ind*-secure and $(n - 2)$*CAKE* secure encryption mode. In the following section, we show how to efficiently compute $y' \cdot A$ by means of bitwise XOR operations.

Bastion: Protocol Specification

We now detail the specification of Bastion.

On input a security parameter k, the key generation algorithm of Bastion outputs a key $K \in \{0, 1\}^k$ for the underlying block-cipher. Bastion leverages block cipher encryption in the CTR mode, which on input a plaintext bitstream x, divides it in blocks $x[1], \ldots, x[m]$, where m is odd[2] such that each block has size l.[3] The set of input blocks is encrypted under key K, resulting in ciphertext $y' = y'[1], \ldots, y'[m + 1]$, where $y'[m + 1]$ is an initialization vector which is randomly chosen from $\{0, 1\}^l$.

Next, Bastion applies a linear transform to y' as follows. Let $n = m + 1$ and assume A to be an n- by- n matrix where element $a_{i,j} = 0^l$ if $i = j$ or $a_{i,j} = 1^l$, otherwise.[4] Bastion computes $y = y' \cdot A$, where additions and multiplications are implemented by means of XOR and AND operations, respectively. That is, $y[i] \in y$ is computed as $y[i] = \bigoplus_{j=1}^{j=n} (y[j] \wedge a_{j,i})$, for $i = 1 \ldots, n$.

Given key K, inverting Bastion entails computing $y' = y \cdot A^{-1}$ and decrypting y' using K. Notice that matrix A is invertible and $A = A^{-1}$. The pseudocode of the encryption and decryption algorithms of Bastion are shown in Algorithms 1 and 2, respectively. Both algorithms use F to denote a generic block cipher (e.g., AES).

In our implementation, we efficiently compute the linear transform using $2n$ XOR operations as follows:

$$t = y'[1] \oplus y'[2] \oplus \cdots \oplus y'[n],\ y[i] = t \oplus y'[i],\ 1 \le i \le n.$$

Note that $y'[1] \ldots y'[n]$ (computed up to line 6 in Algorithm 1) are the outputs of the CTR encryption mode, where $y'[n]$ is the initialization vector. Similar to the CTR encryption mode, the final output of Bastion is one block larger than the original input.

Correctness Analysis

We show that for every $x \in \{0, 1\}^{lm}$ where m is odd, and for every $K \in \{0, 1\}^l$, we have $x = Dec(K, Enc(K,x))$.

In particular, notice that lines 2-6 of Algorithm 1 and lines 9-12 of Algorithm 2 correspond to the standard CTR encryption and decryption routines, respectively.

1. This requirement is essential for the correctness of the sub- sequent linear transform on the ciphertext blocks. That is, if m is even, then the transform is not invertible.
2. l is the block size of the particular block cipher used.
3. 0^l and 1^l denote a bitstring of l zeros and a bitstream of l ones, respectively.

ISBN 978-93-88808-61-3

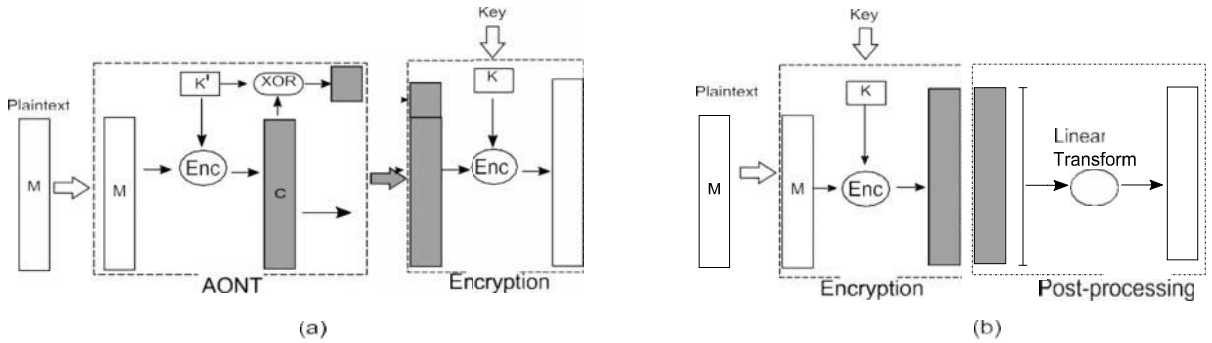


Fig. 2. (a) Current AON encryption schemes require a pre-processing round of block cipher encryption for the AONT, followed by another round of block cipher encryption. (b) On the other hand, Bastion first encrypts the data with one round of block cipher encryption, and then applies an efficient linear post-processing to the ciphertext.

Algorithm 1 Encryption in Bastion.

```
1: procedure Enc(K,x = x[1]...x[m])
2:     n = m + 1
3:     y'[n] ← {0, 1}^l                          ◁ y'[n] is the IV for CTR
4:     for i = 1 ... n − 1 do
5:         y'[i] = x[i] ⊕ F_K(y'[n]+i)
6:     end for
7:     t = 0^l
8:     for i = 1 ... n do
9:         t = t ⊕ y'[i]
10:    end for
11:    for i = 1 ... n do
12:        y[i] = y'[i] ⊕ t
13:    end for
14:    return y                                  ◁ y = y[1] ... y[n]
15: end procedure
```

Algorithm 2 Decryption in Bastion.

```
1: procedure Dec(K, y = y[1] ... y[n])
2: t = 0^l
3: for i = 1 ... n do
4:     t = t ⊕ y[i]
5: end for
6: for i = 1 ... n do
7:     y'[i] = y[i] ⊕ t
8: end for
9: for i = 1 ... n − 1 do
10:    x[i] = y'[i] ⊕ F_K(y'[n] + i)
11: end for
12:    return x                                  ◁ x = x[1] ... x[n − 1]
13: end procedure
```

Therefore, we are only left to show that the linear transformation computed in lines 7-14 of Algorithm 1 is correctly reverted in lines 2-8 of Algorithm 2. In other words, we need to show that $t = \bigoplus_{i=1..n} y[i]$ (as computed in the decryption algorithm) matches $t = \bigoplus_{i=1..n} y[i]'$ (as computed in the encryption algo- rithm).

Recall that t can be computed as follows:

$$
\begin{aligned}
t &= \bigoplus_{i=1..n} y[i] \\
&= \bigoplus_{i=1..n} (y'[i] \oplus t) \\
&= \bigoplus_{i=1..n} y'[i] \oplus \bigoplus_{i=1..n} y'[i] \\
&= \bigoplus_{i=1..n} \bigoplus_{j=1..n, j \neq i} y'[j] \\
&= \bigoplus_{i=1..n} y'[i]
\end{aligned}
$$

Notice that the last step holds because n is even and therefore each $y'[j]$ is XORed for an odd number of times.

REMARK 5. We point out that Bastion is not restricted to the CTR encryption mode and can be instanti- ated with other ind-secure block cipher (and stream ciphers) modes of encryption (e.g., CBC, OFB).

Tointerface with our cloud storage model described in Section 3.1, we assume that each user encrypts the data using Bastion before invoking the write() routine. More specifically, let $\mathrm{Enc}(K, \cdot)$, $\mathrm{Dec}(K, \cdot)$ denote the encryption and decryption routines of Bastion, respectively. Given encryption key K and a file f, the user computes $v \leftarrow \mathrm{Enc}(K, f)$ and invokes write(v) in order to upload the encrypted file to the cloud. In this setting, key K remains stored at the user'smachine. Similarly, to download the file from the cloud, the user invokes read($\cdot$) to fetch v and runs $f \leftarrow \mathrm{Dec}(K, v)$ to recover f.

Security Analysis

In this section, we show that Bastion is *mathrmind* secure and $(n - 2)$*CAKE* secure.

LEMMA 1. Bastion is ind secure.

Proof 1. Bastion uses an *ind* secure encryption mode to encrypt a message, and then applies a linear

ISBN 978-93-88808-61-3

transform on the ciphertext blocks. It is straight- forward to conclude that Bastion is *ind* secure. In other words, a polynomial-time algorithm A that has non-negligible advantage in breaking the *ind* security of Bastion can be used as a black-box by another polynomial-time algorithm B to break the *ind* security of the underlying encryption mode. In particular, B forwards A's queries to its oracle and applies the linear transformation of Algorithm 1 lines 7-14 to the received ciphertext before forward- ing it to A. The same strategy is used when A outputs two messages at the end of the *find* stage: the two messages are forwarded to B's oracle; upon receiving the challenge ciphertext, B applies the linear transformation and forwards it to A. When A replies with its guess b', B outputs the same guess. It is easy to see that if A has non-negligible advantage in guessing correctly which message was encrypted, so does B. Furthermore, the running time of B is the one of A plus the time to apply the linear transformation to A's queries.

LEMMA 2. Given any $n-2$ blocks of $y[1] \ldots y[n]$ as output by Bastion, it is infeasible to compute any $y'[i]$, for $1 \leq i \leq n$.

Proof 2. Let $y = y[1], \ldots, y[n] \leftarrow E(K, x = x[1] \ldots x[m])$. Note that given any $(n-1)$ blocks of y, the adversary can compute one block of y'. In particular, $y'[i] = \bigoplus_{j=1, j \neq i}^{n} y[j]$, for any $1 \leq i \leq n$. As it will become clear later, with one block $y'[i]$ and the encryption key, the adversary has non-negligible probability of winning the game of Definition 3. However, if only $(n-2)$ blocks of y are given, then each of the n blocks of y' can take on any possible values in $\{0,1\}^l$, depending on the two unknown blocks of y. Recall that each block $y'[i]$ is dependent on $(n-1)$ blocks of y and it is pseudo-random as output by the CTR encryption mode. Therefore, given any $(n-2)$ blocks of y, then $y'[i]$ could take any of the 2^l possibilities, for $1 \leq i \leq n$.

LEMMA 3. Bastion is $(n-2)$*CAKE* secure.

Proof 3. The security proof of Bastion resembles the standard security proof of the CTR encryption mode and relies on the existence of pseudo-random permutations. In particular, given a polynomial-type algorithm A which has non-negligible advantage in the $(n-\lambda)$*CAKE* experiment with $\lambda = 2$, we can construct a polynomial-time algorithm B which has non-negligible advantage in distinguishing between a true random permutation and a pseudo-random permutation.

B has access to oracle O and uses it to answer the encryption and decryption queries issued by A. In particular, A's queries are answered as follows:

- *Decryption query for* $y[1] \ldots y[n]$
  1) Compute $t = y[1] \oplus \ldots \oplus y[n]$
  2) Compute $y'[i] = y[i] \oplus t$, for $1 \leq i \leq n$
  3) Compute $x[i] = y'[i] \oplus O(y'[n]+i)$, for $1 \leq i \leq n-1$
  4) Return $x[1] \ldots x[n-1]$
- *Encryption query for* $x[1] \ldots x[n-1]$
  1) Pick random $y'[n] \in \{0,1\}^l$
  2) Compute $y'[i] = x[i] \oplus O(y'[n]+i)$, for $1 \leq i \leq n-1$
  3) Compute $t = y'[1] \oplus \ldots \oplus y'[n]$
  4) Compute $y[i] = y'[i] \oplus t$, for $1 \leq i \leq n$
  5) Return $y[1] \ldots y[n]$

When A outputs two messages $x_1[1] \ldots x_1[n-1]$ and $x_2[1] \ldots x_2[n-1]$, B picks $b \in \{0,1\}$ at random and does the following:

1) Pick random $y'_b[n] \in \{0,1\}^l$
2) Compute $y'_b[i] = x_b[i] \oplus O(y'_b[n], i)$, for $1 \leq i \leq n-1$
3) Compute $t = y_b[1] \oplus \ldots \oplus y_b[n]$
4) Compute $y_b[i] = y'_b[i] \oplus t$, for $1 \leq i \leq n$

At this point, A selects $(n-2)$ indexes $i_1, \ldots i_{n-2}$ and B returns the corresponding $y_b[i_1], \ldots, y_b[i_{n-2}]$. Encryption and decryption queries are answered as above. When A outputs its answer b', B outputs 1 if $b = b'$, and 0 otherwise. It is straightforward to see that if A has advantage larger than negligible to guess b, then B has advantage larger than negligible to distinguish a true random permutation from a pseudorandom one. Furthermore, the number of queries issued by B to its oracle amounts to the number of encryption and decryption queriesissued by A. Note that by Lemma 2, during the guess stage, A cannot issue a decryption query on the challenge ciphertext since with only $(n-2)$ blocks, finding the remaining blocks is infeasible.

REMARK 6. Bastion is not $(n-1)$*CAKE* secure. As shown in the proof of Lemma 2, the adversary can recover one block of y' given any $(n-1)$ blocks of y. If the adversary recovers $y'[n]$ that is used as an IV in the CTR encryption mode, the adversary can easily win the $(n-1)$*CAKE* game. Recall that our security definition allows the adversary to learn the encryption key.

REMARK 7. Bastion is $(n-2)$*CAKE* secure according to Definition 3. However, in a practical deployment, we expect that each file spans several thousands blocks [5]. When those blocks are evenly spread across servers, each server will store a larger number of blocks. Therefore, an $(n-2)$*CAKE* secure scheme such as Bastion clearly preserves data confidential- ity unless *all* servers are compromised.

4. For example, a 10MB file encrypted using AES has more than 600K blocks.

TABLE 1
Comparison between Bastion and existing constructs. We assume a plaintext of m = n − 1 blocks. Since all schemes are symmetric, we only show the computation overhead for the encryption/encoding routine in the column "Computation" ("b.c." is the number of block cipher operations; "XOR" is the number of XOR operations.

	Computation	**Storage** (blocks)	**Security**
CTR Encryption	n − 1 b.c. n − 1 XOR	n	1*CAKE* ind-secure
Rivest AONT [26]	2(n − 1) b.c. 3(n − 1) XOR	n	N/A ind-INsecure
Desai AONT [12]	n − 1 b.c. 2(n − 1) XOR	n	N/A ind-INsecure
Rivest AON Encryption [26]	3n − 2 b.c. 3(n − 1) XOR	n	(n − 1)*CAKE* ind-secure
Desai AON Encryption [12]	2n − 1 b.c. 2(n − 1) XOR	n	(n − 1)*CAKE* ind-secure
Encrypt-then -secret-share	n − 1 b.c. 2n − 1 XOR	n^2	(n−1)*CAKE* ind-INsecure*
Bastion	n − 1 b.c. 3n − 1 XOR	n	(n − 2)*CAKE* ind-secure

* Recall that an *ind*-adversary can access all storage servers to fetch all ciphertext blocks. Therefore, the adversary can also fetch all the key shares and compute the encryption key.

5 Comparison to Existing Schemes

In what follows, we briefly overview several encryption modes and argue about their security (according to Definitions 1 and 3) and performance when compared to Bastion.

CPA-encryption modes

Traditional CPA-encryption modes, such as the CTR mode, provide *ind* security but are only 1*CAKE* secure. That is, an adversary equipped with the encryption key must only fetch two ciphertext blocks to break data confidentiality.[6]

CPA-encryption and secret-sharing

Another option is to rely on the combination of CPA secure encryption modes and secret-sharing.

If the file f is encrypted and then shared with an n-out-of-n secret-sharing scheme(denoted as "encrypt- then-secret-share" in the following), then the construc- tion is clearly $(n-1)$*CAKE* secure and is also *ind* secure. However, secret-sharing the ciphertext comes at considerable storage costs; for example, each share would be as large as the file f using a perfect secret sharing scheme—which makes it impractical for storing large files.

Secret-sharing the encryption key and dispersing its shares across the storage servers alongside the cipher- text is not secure against an *ind*-adversary. Indeed, if the adversary can access all the storage servers and down- load all ciphertext blocks, the adversary may as well download all key shares and compute the encryption key.

1. We assume that the CTR encryption routine starts with a random IV that is incremented at every block encryption.

AON encryption

Recall that an AONT is not an encryption scheme and does not require the decryptor to have any secret key. That is, an AONT is not secure against an *ind*-adversary which can access all the ciphertext blocks. One alter- native is to combine the use of AONT with standard encryption. Rivest [26] suggests to pre-process a mes- sage with an AONT and then encrypt its output with an encryption mode. This paradigm is referred to in the literature as AON encryption and provides $(n-1)$*CAKE* security. Existing AON encryption schemes require at least two rounds of block cipher encryption with two different keys [12], [26]. At least one round is required for the actual AONT that embeds the first encryption key in the pseudo-ciphertext (cf. Section 2). An addi- tional round uses another encryption key that is kept secret to guarantee CPA-security. However, two encryp- tion rounds constitute a considerable overhead when encrypting and decrypting large files. In Appendix A, we describe possible ways of modifying the AONTs of [26] and [12] to achieve *ind* security and $(n-1)$*CAKE* security without adding another round of block cipher encryption, and we discuss their shortcomings.

Clearly, these solutions are either not satisfactory in terms of security or incur a large overhead when compared to Bastion and may not be suitable to store large files in a multi-cloud storage system.

Performance Comparison

Table 1 compares the performance of Bastion with the encryption schemes considered so far, in terms of computation, storage, and security.

Given a plaintext of m blocks, the CTR encryption mode outputs $n = m + 1$ ciphertext blocks, computed with $(n-1)$ block cipher operations and $(n-1)$ XOR

operations. The CTR encryption mode is *ind* secure but only 1*CAKE* secure.

Rivest AONT outputs a pseudo-ciphertext of $n = m + 1$ blocks using $2(n - 1)$ block cipher operations and $3(n-1)$ XOR operations. Desai AONT outputs the same number of blocks but requires only $(n - 1)$ block cipher operations and $2(n - 1)$ XOR operations. Both Rivest AONT and Desai AONT are, however, not *ind* secure since the encryption key used to compute the AONT output is embedded in the output itself. Encrypting the output of Rivest AONT or Desai AONT with a stan- dard encryption mode (both [12] and [26] use the ECB encryption mode), requires additional n block cipher operations, and yields an AON encryption that is *ind* secure[7] and $(n - 1)$*CAKE* secure. Encrypt-then-secret- share (cf. Section 4.4) is *ind* secure and $(n - 1)$*CAKE* secure. It requires $(n - 1)$ block cipher operations and n XOR operations if additive secret sharing is used. How- ever secret-sharing encryption results in a prohibitively large storage overhead of n^2 blocks.

Bastion also outputs $n = m + 1$ ciphertext blocks. It achieves *ind* security and $(n - 2)$*CAKE* security with only $(n - 1)$ block cipher operations and $(3n - 1)$ XOR operations.[8]

We conclude that Bastion achieves a solid tradeoff between the computational overhead of existing AON encryption modes and the exponential storage overhead of secret-sharing techniques, while offering a compa- rable level of security. In Section 6, we confirm the superior performance of Bastion by means of imple- mentation.

6 Implementation and Evaluation

In this section, we describe and evaluate a prototype implementation modeling a read-write storage system based on Bastion. We also discuss insights with respect to the integration of Bastion within existing dispersed storage systems.

Implementation Setup

Our prototype, implemented in C++, emulates the read- write storage model of Section 3.1. We instantiate Bastion with the CTR encryption mode (cf. Figure 1) using both AES128 and Rijndael256, implemented using the libmcrypt.so. 4.4.7 library. Since this library doesnot natively support the CTR encryption mode, we use it for the generation of the CTR keystream, which is later XORed with the plaintext.

We compare Bastion with the AON encryption schemes of Rivest [26] and Desai [12]. For baseline comparison, we include in our evaluation the CTR encryption mode and the AONTs due to Rivest [26] and Desai [12], which are used in existing dispersed storage systems, e.g., Cleversafe [25]. We do not evaluate the performance of secret-sharing the data because of its prohibitively large storage overhead (squared in the number of input blocks). We evaluate our implemen- tations on an Intel(R) Xeon(R) CPU E5-2470 running at 2.30GHz. Note that the processor clock frequency might have been higher during the evaluation due to the TurboBoost technology of the CPU. In our evaluation, we abstract away the effects of network delays and congestion, and we only assess the processing perfor- mance of the encryption for the considered schemes. This is a reasonable assumption since all schemes are length-preserving (plus an additional block of l bits), and are therefore likely to exhibit the same network performance. Moreover, we only measure the per- formance incurred during encryption/encoding, since all schemes are symmetric, and therefore the decryp- tion/decoding performance is comparable to that of the encryption/encoding process.

We measure the peak throughput and the latency ex- hibited by our implementations w.r.t. various file/block sizes. For each data point, we report the average of 30 runs. Due to their small widths, we do not show the corresponding 95% confidence intervals.

Evaluation Results

Our evaluation results are reported in Figure 3 and Figure 4. Both figures show that Bastion considerably improves (by more than 50%) the performance of ex- isting $(n - 1)$*CAKE* encryption schemes and only in- curs a negligible overhead when compared to existing semantically secure encryption modes (e.g., the CTR encryption mode) that are only 1*CAKE* secure.

In Figure 3, we show the peak throughput achie- ved by the CTR encryption mode, Bastion, Desai AONT/AON, and Rivest AONT/AON schemes. The peak throughput achieved by Bastion reaches almost 72 MB/s and is only 1% lower than the one exhibited by the CTR encryption mode. When compared with ex- isting $(n - 1)$*CAKE* secure schemes, such as Desai AON encryption and Rivest AON encryption, our results show that the peak throughput of Bastion is almost twice as large as that of Desai AON encryption, and more than three times larger than the peak throughput of Rivest AON encryption.

We also evaluate the performance of Bastion, with respect to different block sizes of the underlying block cipher. Our results show that—irrespective of the block size—Bastion only incurs a negligible performance de- terioration in peak throughput when compared to the CTR encryption mode. Figures 4(a) and 4(b) show the latency (in ms) incurred by the encryption/encoding routines for different file sizes. The latency of Bastion is comparable to that of the CTR encryption mode—for both AES128 and Rijandael256—and results in a con- siderable improvement over existing AON encryption schemes (more than 50% gain in latency).

1. Security according to Definition 1 is achieved because the key used to create the AONT is always random, even if the key used to add the outer layer of encryption is fixed.

2. Bastion requires (n− 1) XOR operations for the CTR encryp- tion and 2n XOR operations for the linear transform.

ISBN 978-93-88808-61-3

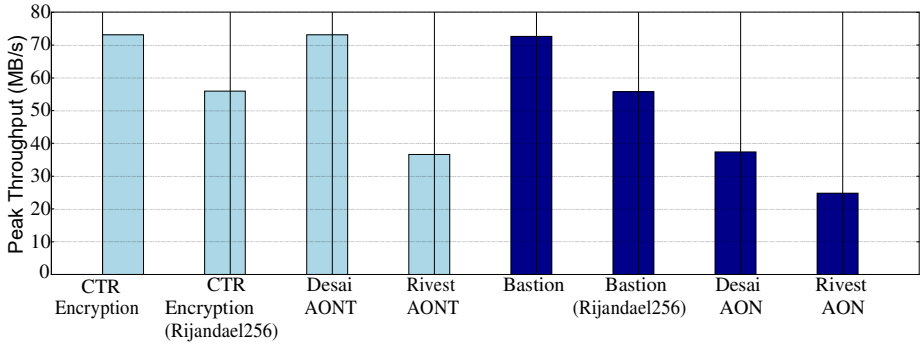
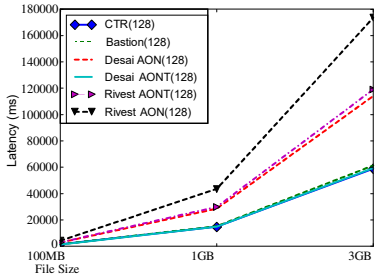
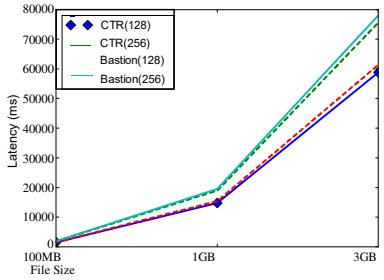


Fig. 3. Peak throughput comparison. Unless otherwise specified, the underlying block cipher is AES128. Each data point is averaged over 30 runs. Histograms in dark blue depict encryption modes which offer comparable security to Bastion. Light blue histograms refer to encryption/encoding modes where individual ciphertext blocks can be inverted when the key is exposed.



(a) Latency of encryption/encoding for differ- ent file sizes.



(b) Latency of encryption/encoding for different block sizes of the underlying block cipher.

Fig. 4. Performance evaluation of Bastion. Each data point in is averaged over 30 runs. Unless otherwisespecified, the underlying block cipher is AES-128. CTR(256) and Bastion(256) denote the CTR encryption mode and Bastion encryption routine, respectively, instantiated with Rijandael256.

Deployment within HYDRAstor

Recall that Bastion preserves data confidentiality against an adversary that has the encryption key as long as the adversary does not have access to two ciphertext blocks. In a multi-cloud storage system, if each server stores at least two ciphertext blocks, then Bastion clearly preserves data confidentiality unless *all* servers are compromised.

In scenarios where servers can be faulty, Bastion can be combined with information dispersal algorithms (e.g., [24]) to provide data confidentiality and fault tolerance. Recall that information dispersal algorithms (IDA), parameterized with t_1, t_2 (where $t_1 \leq t_2$), encode data into t_2 symbols such that the original data can be recovered from any t_1 encoded symbols. In our multi- cloud storage system (cf. Section 3.1), the ciphertext output by Bastion is then fed to the IDA encoding routine, with symbols of size l bits, and with parameters $t_2 \geq 2s$, $t_1 < t_2$, where s is the number of available servers. Since the output of the IDA is equally spread across the s servers, by setting $t_2 \geq 2s$, we ensure that each server stores at least two ciphertext blocks worth of data. Finally, the encoded symbols are input to the write() routine that distributes symbols evenly to each of the storage servers. Recovering f via the read() routine entails fetching t_1 encoded symbols from the servers and decoding them via the IDA decoding routine. The resulting ciphertext can be decrypted using Bastion to recover file f. By doing so, data confiden- tiality is preserved even if the key is exposed unless

$t = \frac{st_1}{t_2}$ servers are compromised. Furthermore, data availability is guaranteed in spite of $(s - t)$ server failures.

HYDRAstor

We now discuss the integration of a prototype implementation of Bastion within the HYDRAstor grid storage system [13], [23]. HYDRAstor is a commer- cial secondary storage solution for enterprises, which consists of a back-end architectured as a grid of stor- age nodes built around a distributed hash table. HY- DRAstor tolerates multiple disk, node and network failures, rebuilds the data automatically after failures, and informs users about recoverability of the deposited data [13]. The reliability and availability of the stored data can be dynamically adjusted by the clients with each write operation, as the back-end supports multiple data resiliency classes [13].

HYDRAstor distributes written data to multiple disks using the distributed resilient data technology (DRD); the combination of Bastion with DRD ensures that an adversary which has the encryption key and compromises a subset of the disks (i.e., determined by the reconstruction threshold), cannot acquire any mean- ingful information about the data stored on the disk. To better assess the performance impact of Bastion in HYDRAstor, we evaluated the performance of Bastion in the newest generation HYDRAstor HS8-4000 series system, which uses CPUs with accelerated AES encryption (i.e., the AESNI instruction set). In our experiments, all written data was unique to remove the effect of data deduplication. Results show that the write bandwidth was not affected by the integration of Bastion. The read bandwidth decreased only by 3%. In both read and write operations, the CPU utilization in the system only increased marginally. These experiments clearly suggest that Bastion can be integrated in existing com- mercial storage systems to strengthen the security of these systems under key exposure, without affecting performance.

7 Related Work

To the best of our knowledge, this is the first work that addresses the problem of securing data stored in multi- cloud storage systems when the cryptographic material is exposed. In the following, we survey relevant related work in the areas of deniable encryption, information dispersal, all-or-nothing transformations, secret-sharing techniques, and leakage-resilient cryptography.

Deniable Encryption

Our work shares similarities with the notion of "shared- key deniable encryption" [9], [14], [18]. An encryption scheme is "deniable" if—when coerced to reveal the en- cryption key—the legitimate owner reveals "fake keys" thus forcing the ciphertext to "look like" the encryption of a plaintext different from the original one—hence keeping the original plaintext private. Deniable en- cryption therefore aims to deceive an adversary which does not know the "original" encryption key but, e.g., can only acquire "fake" keys. Our security definition models an adversary that has access to the real keying material.

Information Dispersal

Information dispersal based on erasure codes [30] has been proven as an effective tool to provide reliability in a number of cloud-based storage systems [1], [2], [20], [33]. Erasure codes enable users to distribute their data on a number of servers and recover it despite some servers failures.

Ramp schemes [7] constitute a trade-off between the security guarantees of secret sharing and the efficiency of information dispersal algorithms. A ramp scheme achieves higher "code rates" than secret sharing and features two thresholds t_1, t_2. At least t_2 shares are required to reconstruct the secret and less than t_1 shares provide no information about the secret; a number of shares between t_1 and t_2 leak "some" information.

All or Nothing Transformations

All-or-nothing transformations (AONTs) were first introduced in [26] and later studied in [8], [12]. The majority of AONTs leverage a secret key that is em- bedded in the output blocks. Once all output blocks are available, the key can be recovered and single blocks can be inverted. AONT, therefore, is not an encryption scheme and does not require the decryptor to have any key material. Resch et al. [25] combine AONT and information dispersal to provide both fault-tolerance and data secrecy, in the context of distributed storage systems. In [25], however, an adversary which knows the encryption key can decrypt data stored on single servers.

Secret Sharing

Secret sharing schemes [5] allow a dealer to distribute a secret among a number of shareholders, such that only authorized subsets of shareholders can reconstruct the secret. Inthresholdsecret sharing schemes[11], [27], the dealer defines a threshold t and each set of shareholders of cardinality equal to or greater than t is authorized to reconstruct the secret. Secret sharing guarantees se- curity against a non-authorized subset of shareholders; however, they incur a high computation/storage cost, which makes them impractical for sharing large files. Rabin [24] proposed an information dispersal algorithm with smaller overhead than the one of [27], however the proposal in [24] does not provide any security guarantees when a small number of shares (less than the reconstruction threshold) are available. Krawczyk
[19] proposed to combine both Shamir's [27] and Ra- bin's [24] approaches; in [19] a file is first encrypted using AES and then dispersed using the scheme in [24], while the encryption key is shared using the scheme in [27]. In Krawczyk's scheme, individual ciphertext blocks encrypted with AES can be decrypted once the key is exposed.

Leakage-resilient Cryptography

Leakage-resilient cryptography aims at designing cryptographic primitives that can resist an adversary which learns partial information about the secret state of a sys- tem, e.g., through side-channels [22]. Different models allow to reason about the "leaks" of real implemen- tations of cryptographic primitives [22]. All of these models, however, limit in some way the knowledge of the secret state of a system by the adversary. In contrast, the adversary is given all the secret material in our model.

8 CONCLUSION

In this paper, we addressed the problem of securing data outsourced to the cloud against an adversary which has access to the encryption key. For that pur- pose, we introduced a novel security definition that captures data confidentiality against the new adversary. We then proposed Bastion, a scheme which ensures the confidentiality of encrypted data even when the adversary has the encryption key, and all but *two* cipher- text blocks. Bastion is most suitable for settings where the ciphertext blocks are stored in multi-cloud storage systems. In these settings, the adversary would need to acquire the encryption key, and to compromise *all* servers, in order to recover any single block of plaintext.

We analyzed the security of Bastion and evaluated its performance in realistic settings. Bastion consider- ably improves (by more than 50%) the performance of existing primitives which offer comparable secu- rity under key exposure, and only incurs a negligible overhead (less than 5%) when compared to existing semantically secure encryption modes (e.g., the CTR encryption mode). Finally, we showed how Bastion can be practically integrated within existing dispersed storage systems.

REFERENCES

[1] M. Abd-El-Malek, G. R. Ganger, G. R. Goodson, M. K. Re- iter, and J. J. Wylie, "Fault-Scalable Byzantine Fault-Tolerant Services," in *ACM Symposium on Operating Systems Principles (SOSP)*, 2005, pp. 59–74.

[2] M. K. Aguilera, R. Janakiraman, and L. Xu, "Using Erasure Codes Efficiently for Storage in a Distributed System," in *International Conference on Dependable Systems and Networks (DSN)*, 2005, pp. 336–345.

[3] W. Aiello, M. Bellare, G. D. Crescenzo, and R. Venkatesan, "Security amplification by composition: The case of doubly-iterated, ideal ciphers," in *Advances in Cryptology (CRYPTO)*, 1998, pp. 390–407.

[4] C. Basescu, C. Cachin, I. Eyal, R. Haas, and M. Vukolic, "Ro- bust Data Sharing with Key-value Stores," in *ACM SIGACT- SIGOPS Symposium on Principles of Distributed Computing (PODC)*, 2011, pp. 221–222.

[5] A. Beimel, "Secret-sharing schemes: A survey," in *Interna- tional Workshop on Coding and Cryptology (IWCC)*, 2011, pp. 11–46.

[6] A. Bessani, M. Correia, B. Quaresma, F. André, and P. Sousa, "DepSky: Dependable and Secure Storage in a Cloud-of- clouds," in *Sixth Conference on Computer Systems (EuroSys)*, 2011, pp. 31–46.

[7] G. R. Blakley and C. Meadows, "Security of ramp schemes," in *Advances in Cryptology (CRYPTO)*, 1984, pp. 242–268.

[8] V. Boyko, "On the Security Properties of OAEP as an All- or-nothing Transform," in *Advances in Cryptology (CRYPTO)*, 1999, pp. 503–518.

[9] R. Canetti, C. Dwork, M. Naor, and R. Ostrovsky, "Deniable Encryption," in *Proceedings of CRYPTO*, 1997.

[10] Cavalry, "Encryption Engine Dongle," http://www. cavalrystorage.com/en2010.aspx/.

[11] C. Charnes, J. Pieprzyk, and R. Safavi-Naini, "Conditionally secure secret sharing schemes with disenrollment capability," in *ACM Conference on Computer and Communications Security (CCS)*, 1994, pp. 89–95.

[12] A. Desai, "The security of all-or-nothing encryption: Protect- ing against exhaustive key search," in *Advances in Cryptology (CRYPTO)*, 2000, pp. 359–375.

[13] C. Dubnicki, L. Gryz, L. Heldt, M. Kaczmarczyk, W. Kil- ian, P. Strzelczak, J. Szczepkowski, C. Ungureanu, and M. Welnicki, "HYDRAstor: a Scalable Secondary Storage," in *USENIX Conference on File and Storage Technologies (FAST)*, 2009, pp. 197–210.

[14] M. Dürmuth and D. M. Freeman, "Deniable encryption with negligible detection probability: An interactive construction," in *EUROCRYPT*, 2011, pp. 610–626.

[15] EMC, "Transform to a Hybrid Cloud," http://www.emc. com/campaign/global/hybridcloud/index.htm.

[16] IBM, "IBM Hybrid Cloud Solution," http://www-01.ibm. com/software/tivoli/products/hybrid-cloud/.

[17] J. Kilian and P. Rogaway, "How to protect DES against exhaustive key search," in *Advances in Cryptology (CRYPTO)*, 1996, pp. 252–267.

[18] M. Klonowski, P. Kubiak, and M. Kutylowski, "Practical De-niable Encryption," in *Theory and Practice of Computer Science (SOFSEM)*, 2008, pp. 599–609.

[19] H. Krawczyk, "Secret Sharing Made Short," in *Advances in Cryptology (CRYPTO)*, 1993, pp. 136–146.

[20] J. Kubiatowicz, D. Bindel, Y. Chen, S. E. Czerwinski, P. R. Eaton, D. Geels, R. Gummadi, S. C. Rhea, H. Weatherspoon, W. Weimer, C. Wells, and B. Y. Zhao, "OceanStore: An Archi-tecture for Global-Scale Persistent Storage," in *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2000, pp. 190–201.

[21] L. Lamport, "On interprocess communication," 1985.

[22] S. Micali and L. Reyzin, "Physically observable cryptography (extended abstract)," in *Theory of Cryptography Conference (TCC)*, 2004, pp. 278–296.

[23] NEC Corp., "HYDRAstor Grid Storage," http://www. hydrastor.com.

[24] M. O. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance," *J. ACM*, vol. 36, no. 2, pp. 335–348, 1989.

[25] J. K. Resch and J. S. Plank, "AONT-RS: Blending Security and Performance in Dispersed Storage Systems," in *USENIX Conference on File and Storage Technologies (FAST)*, 2011, pp. 191–202.

[26] R. L. Rivest, "All-or-Nothing Encryption and the Package Transform," in *International Workshop on Fast Software Encryp-tion (FSE)*, 1997, pp. 210–218.

[27] A. Shamir, "How to Share a Secret?" in *Communications of the ACM*, 1979, pp. 612–613.

[28] D. R. Stinson, "Something About All or Nothing (Trans- forms)," in *Designs, Codes and Cryptography*, 2001, pp. 133– 138.

[29] StorSimple, "Cloud Storage," http://www.storsimple.com/.

[30] J. H. van Lint, *Introduction to Coding Theory*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1982.

[31] Wikipedia, "Edward Snowden," http://en.wikipedia.org/ wiki/Edward_Snowden#Disclosure.

[32] Z. Wu, M. Butkiewicz, D. Perkins, E. Katz-Bassett, and H. V. Madhyastha, "SPANStore: Cost-effective Geo-replicated Stor- age Spanning Multiple Cloud Services," in *ACM Symposium on Operating Systems Principles (SOSP)*, 2013, pp. 292–308.

[33] H. Xia and A. A. Chien, "RobuSTore: a Distributed Stor- age Architecture with Robust and High Performance," in *ACM/IEEE Conference on High Performance Networking and Computing (SC)*, 2007, p. 44.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Data Implementation on Friendly Environment with Green Data Mining Process

Tabitha indupalli
tabi.indupalli@gmail.com

Sharon Drisilda
emmanuelalfred14@gmail.com

Abstract

This paper develops a set of principles for green data mining, related to the key stages of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The principles are grounded in a review of the Cross Industry Standard Process for Data mining (CRISP-DM) model and relevant literature on data mining methods and Green IT. We describe how data scientists can contribute to designing environmentally friendly data mining processes, for instance, by using green energy, choosing between make-or-buy, exploiting approaches to data reduction based on business understanding or pure statistics, or choosing energy friendly models.

1. Introduction

The use of computing power coupled with the unprecedented availability of data provide ample opportunity to improve energy efficiency . However, they are also an increasingly relevant source of energy consumption and associated carbon emissions. Data centers consumed about 70 billion kWh in 2016 in the United States alone , and the total consumption of all IT is estimated to be close to 5% of total energy consumption. In response to this increasing amount of energy used by IT, Greenpeace published the "Guide to Building the Green Internet" , promoting "a more widespread adaption in best practices" for energy efficient data center design. They demand that "data center operators and customers should regularly report their energy performance and establish transparent energy savings targets." Electricity consumption is costly—it involves various detrimental effects on nature and society, ranging from bird deaths by wind turbines, on to severe air pollution and CO2 emissions by coal power plants, and the risk of catastrophes stemming from nuclear power plants.

These concerns are partially addressed by current initiatives under notions such as green information systems (Green IS) or green information technology (Green IT) , but environmentally friendly data mining is a novel topic.

Data scientists often leverage a large pool of computational resources using sophisticated and computationally costly machine learning techniques to extract knowledge and insights from data. Though existing processes such as the Cross Industry Standard Process for Data mining (CRISP-DM) provide some guidance on how to execute a data mining project, the skills of a data scientist heavily rely on creativity involving many degrees of freedom, often including the choice of tools, models, and data sources.

It is against this background that, in this paper, we develop guidelines for data scientists to implement more environmentally friendly practices that can complement technology-focused perspectives aiming to design more energy efficient IT-based systems. Specifically, we are focusing attention on one important area of data science—data mining. Data mining can be described as knowledge discovery from data or in terms of different activities as collecting, cleaning, processing, analyzing and gaining useful insights from data . We ask: How can data *scientists implement more environmentally friendly data mining processes?*

The remainder of this paper is structured as follows. We first describe our methodology. We then review the data mining process and develop a set of principles for green data mining. We conclude by discussing limitations and future work.

2. Methodology

We derived our principles by analyzing the CRISP-DM data mining process and literature on green IT and data mining. In a first step, we identified factors determining energy consumption. In a second step, we identified individual steps of the CRISP-DM process by investigating possibilities for reduction of each factor. We limited our analysis to those aspects that can be directly influenced by data scientists, including the choice of data, its representation, as well as processes and techniques used throughout the data

analysis process. We do not target the development of novel data mining algorithms for specific problems or improving hardware or software, though some of our insights might be helpful in guiding such developments.

We conducted a narrative literature review on green IT, green IS, and data mining because our goal was to investigate elementary factors and research outcomes related to these areas of research. Green data science is a novel field and, therefore, is more amenable to a qualitative approach such as narrative literature review than a more quantitative approach detailing the current-state-of-research, as done for a descriptive review. Our focus was on using established online databases from computer science as well as information systems such as IEEE Xplore, ProQuest (ABI/INForm), ScienceDirect (Elsevier), AIS electronic library and the ACM digital library. We did not limit ourselves to journals since new ideas are often presented first at academic conferences and a significant body of works, in particular in the field of computer science, only appear as conference articles.

3. The data mining process

There are multiple data mining processes , most of which share common phases. CRISP-DM is arguably the most widely known and practiced model , attending to business and data understanding, data preparation, modelling, evaluation and deployment (Figure 1). The business understanding phase clarifies project objectives and business requirements, which are then translated into a data mining problem. There are unsupervised data mining problems including association pattern mining and clustering as well as supervised approaches like classification . Data understanding typically requires initial data selection or collection. Data is first analyzed in an exploratory fashion to get a basic understanding of the data in the business context. Exploratory analysis supports the development of hypothesis by identifying patterns in the data [3]. It allows to get first insights as well as to identify data quality problems. Data preparation includes using raw data to derive data that can be fed into the models. Activities include data selection, transformation, and cleaning. The data might have to be prepared separately for each model. The modelling phase consists of defining suitable models, selecting a model, and adapting the model, for instance, optimizing its parameters to solve the data mining problem. Computational evaluation of the model is part of the model selection process. Every data mining problem can be tackled using different strategies and models. Generally, there is no clear consensus about which model is best for a task. Consequently, some form of trial and error can often not be avoided. This is supported by the “no free lunch” theorem stating that any algorithm outperforms any other algorithm on some datasets as well as by empirical studies

. The choice of models depends on many factors such as data (dimensionality, number of observations, structuredness), data mining objectives (need for best possible expected outcome, need to explain results), and cost (focus on minimum human effort to build or operate). From the perspective of green data mining, performance is assessed in terms of energy consumption for model training and model use, for instance, for making predictions. For the evaluation phase the main goal is to review all steps involved in the construction of the model, and to verify whether the final model meets the defined business objectives. If the best model meets the evaluation criteria, then it is deployed. Deployment ranges from fabricating a report presenting the findings in an easy-to-comprehend manner to implementing a long running system. Such a system might learn continuously while often performing a prediction task.

4. Principles of green data mining

Grounded in concepts and ideas from the literature on Green IT as well as data mining and its processes,

Table 1: Factors and methods related to green data mining

Factor	Subfactors	Methods for Green Data Mining
Project Objectives and Execution	Performance specification; Make, buy, share	Transfer Learning
Data	Quantity; Quality; Representation; Data acquisition method; Data storage	Sampling, Active Learning, Dimensionality Reduction, Compression, Change of Data Representation, Data Aggregation
Computation (Analysis)	Structuring of computation; Choice/Training of models; Training of models	Reuse of intermediate results; Approximate Models/Algorithms
IT Infrastructure	Hardware, e.g., CPU, Storage	

we identified factors determining the ecological footprint of data mining and we developed principles for reducing this footprint (Table 1, Figure 1).

Green IT discusses institutional perspectives , the role of users, including their behavior and beliefs when using IT-based systems as well as technical concerns . Topics include computational methods, their implementation in software , hardware components of computers , datacenters, cloud computing , parallel data processing (for big data) , as well as organizational and business aspects such as sustainable value chains, green oriented procurement, and adoption of Green IT . Loeser et al. discussed constructs and practices from Green IT (and IS) with respect to sourcing, operations, disposal, governance and end products.

Current literature on data mining , in particular data mining processes , does not explicitly discuss environmental concerns of data mining but touches upon aspects related to computational efficiency and storage such as data reduction and approximate algorithms.

Next, we describe principles of green data mining related to the different steps of the CRISP-DM process. We first elaborate on those principles that pertain to all stages of the process (principles 1-3 in Figure 1), before we then turn to those which only address specific stages (principles 4-8).

Principle #1: Identify and focus on the most energy consuming phases

To maximize the outcome of time invested into making data mining more environmentally friendly, the focus should be on the most energy consuming factors. This analysis can be performed by investigating the factors listed in Table 1 and analyzing each process step shown in Figure 1. Which process steps and factors dominate energy consumption depends on the goals and particularities of the data mining endeavor. Project objectives such as predictive accuracy or required confidence in the analysis are very likely to have a profound impact on energy consumption, since they often indirectly influence the choice of computational methods and data. For example, recent "deep learning" methods have outperformed other machine learning approaches for multiple classification tasks. A data scientist might turn to deep learning to meet certain project objectives, because it achieves state-of-the-art performance with respect to accuracy but, at the same time, requires lots of data and computation. Data preparation does often only require simple techniques, but it might be dominating in terms of energy consumption if complex computationally expensive methods are needed to extract features from the data that are used in later phases of the process. Deployment might be the dominating step if a system is built for continuous usage with large amounts of data. Still, deployment might contribute very little to the overall energy consumption compared to model selection, if the goal of the data mining project is to derive a report supporting a one-time decision.

Principle #2: Share and re-use data, models, frameworks and skills

A data scientist might control make-or-buy decisions. For example, for marketing purposes, she might choose to acquire data from social media

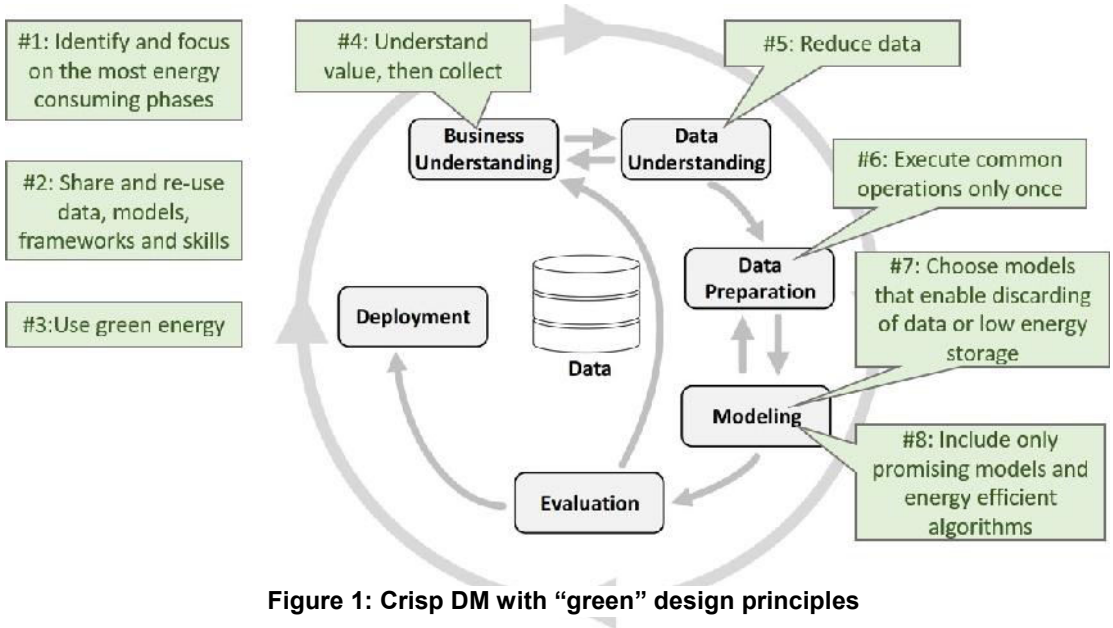


Figure 1: Crisp DM with "green" design principles

channels such as Twitter or Facebook and conduct the analysis by herself. She might also acquire models (implemented in software) to conduct the analysis. She might also decide to consult an external company to conduct the analysis or to obtain models. From an environmental perspective, outsourcing can be preferable if the contractor is more energy-efficient in extracting the demanded information, for instance, because of their prior experience and specialization, more energy efficient infrastructure, or even possession of relevant data. On a global scale, outsourcing of data analysis has the potential to involve less computation and to save energy.

Progress in the field of data science also relies on publicly available data, models, and development frameworks. Initiatives to make data available by research institutions and by governments help create entire ecosystems . State-of-the-art tools to develop (deep learning) models such as Google's Tensorflow are made freely available by large corporations. For such frameworks there are also numerous pre-trained models freely available, e.g., for image recognition based on the Imagenet dataset . Transfer learning is a technique that enables using knowledge from existing models trained for a specific task and dataset on different tasks . The idea is that some "knowledge" of a model can be transferred to another domain. Deep learning networks might benefit from reusing parameters or layers of an already trained network to reduce time (and energy consumption) on developing a new model. Thus, a green data scientist should also contribute data, models, and potentially extensions to frameworks to encourage re-use.

Principle #3: Use green energy

The use of renewable ("green") energy such as solar or wind should be maximized. Conceptually, the idea is to align computation with the availability of green energy. Technical realizations for data processing tasks for distributed data processing platforms (e.g., Hadoop) have been investigated. A system must predict the availability of green energy as well as brown energy and derive a schedule to maximize green energy use and to avoid using brown power at peak demand times. This strategy might also have a positive impact on energy costs as these increase with demand. The data scientist should identify the maximum possible slack in executing data processing tasks based on business objectives. More flexible scheduling allows for using more green energy.

Business understanding

The business understanding phase does typically not involve computation and as such generally does not contribute directly to the energy consumption. Still, understanding the business requirements and trends in the industry sector helps anticipate factors that influence energy consumption of later process steps, such as "What data are relevant and should be collected?" or "What precision of numbers is needed (over time)?" or "How frequently is a deployed system used?" or "How does the value of data change over time?"

Principle #4: Understand value, then collect and forget

Following the idea that "Data is the new oil"—a statement coined by Clive Humbly in 2006—it seems natural to collect as much data as possible, in particular given that storage is cheap and data might generate value "eventually." It is not uncommon that data can be obtained almost for free, for instance, in the form of trace data generated by users visiting a webpage. But, more data increases costs (due to storage and processing), requires more energy, impacts system performance and complexity and, additionally, enhances the risk of information overload. Query times to a database, for instance, increase with the amount of data stored in the database. The idea of collecting data only for the sake of collection has been criticized–"less data can be more value" . The data scientist should thus try to determine what data is relevant for the business or task at hand. Moreover, the quality of the data should be taken into consideration because data of inferior quality might require non-negligible effort for data cleaning .

Not all data has the same value. Even when data consists of a set of observations of the same kind, certain observations might be more valuable than others. For example, for observations, which should be split into classes, "difficult" to classify observations are often more helpful in training data mining models than "easy" to classify observations. Though computational methods can often determine the relevance of data with respect to well-defined metrics, a holistic understanding of the business, its objectives, data, and analytical methodology is essential to limit the collection of data. Leading data analytics companies such as Google embrace the idea of computing on more "little" data, that is, samples . This reasoning is well-founded not only based on statistical models, but also because models benefit from training data in a highly non-linear fashion with decreasing marginal gains given more data . Therefore, in some scenarios, reducing the volume of data might be feasible with considerable impact on energy consumption but only minor changes for other

relevant metrics. Since each model comes with its own strengths and weaknesses related to interpretability, robustness, speed of learning, etc., the overall assessment of advantages and disadvantages must be carefully conducted and aligned with underlying business objectives.

Data understanding

Principle #5: Reduce data

The data scientist might face the choice of what data to collect (or store). This choice must be made with great foresight in order not to miss any opportunity for data-driven value creation. Business understanding as well as an in depth understanding of the data are necessary. However, there are also multiple helpful techniques based on computational and statistical methods that might be supportive. We describe strategies to minimize the amount of data to be collected or used for training such as sampling and dimensionality reduction. These strategies can be employed to limit the number of attributes or observations, reducing precision and changing the representation of data.

Principle #5.1: Reduce number of data items

Often the data scientist can retrieve accurate results by looking at data samples or by using aggregated data. Data can also be categorized (or clustered) into groups, such that different attributes are relevant for some groups but not for others. A group might also be described using an average or median value. The grouping itself might be obtained by clustering algorithms, for instance, documents can be summarized using centroids obtained through clustering. Intuitively, one should maintain data that is most relevant to achieve a certain task. Active learning seeks to incrementally acquire relevant samples for learning. Thus, rather than having a passive model (or learner) that just uses the training data as given, an active learner might ask explicitly for data that is expected to yield maximal improvement in learning. Active learning is typically used in determining what data to collect. But the idea of active learning might also be used to assess the relevance of data and filter data accordingly. A model can be trained using active learning by incrementally adding the most important data items of the full dataset. The learning process might be stopped if there is no more data that improves the model beyond a small threshold. Unused data, which does not improve the model significantly, could then be discarded. Uncertainty sampling is the most prominent technique in active learning in the context of classification . It seeks to obtain labelled data, where there is most uncertainty about the correct class labels. Uncertainty sampling has been employed successfully for margin-based classifiers such as Support Vector Machines (SVMs) . Standard sampling techniques can also be helpful to reduce the amount of data. One of the simplest, but often sufficient approaches is to conduct simple random sampling—choosing each data point with the same probability without replacement of selected data points. In a case study on predicting conversion probabilities for two online retailers, Stange and Funk could show that only 1% of the data available to them was enough to achieve the optimal tradeoff between accuracy and the cost of collecting and processing the data. Stratified sampling is an appropriate sampling technique if groups are homogeneous, that is, data within groups has lower variance than data from distinct groups. One could also employ density-based sampling, for instance, assign samples with lower density a higher probability. This is useful if data from rare regions is highly important.

Principle #5.2: Reduce number or precision of attributes

The dataset might contain attributes that are irrelevant for the analysis. These attributes can be safely neglected. The relevance might depend on the type of data. For many text mining problems very frequent words—so-called stop words, such as “and”, ”the”, “is”, ”are”—can be ignored. In fact, removing unnecessary or noisy attributes such as stop words is often recommended . More generally, dimensionality reduction can be achieved by feature selection and extraction as well as type transformation . Feature selection techniques encompass filter and wrapper methods as well as their combination. Filter models assess the impact of features by some criterion independent of the model. Wrapper models train the model using a subset of features. An example of a filter model is the use of predictive attribute dependence, where the idea is that correlated features yield better outcomes than uncorrelated ones. Therefore, the relevance of an attribute might be determined by assessing the classification accuracy when using all other attributes to predict the attribute. These techniques can be employed to remove attributes that do not reach a minimum relevance threshold. Since many of the techniques are of heuristic nature, the impact of the removal of data that is deemed irrelevant should be tested, for instance, by comparing models being trained on the full and the reduced attribute set. Attribute reduction can also lead to an increase in accuracy, e.g., for decision trees

Feature extraction is often performed through axis rotations in a way that axes are sorted according to their ability to reconstruct data with minimal error.

Axes with negligible impact on data reconstruction can be removed. The derived dataset can often be used to train a model or it might be used to reconstruct the original data, which in turn is used for training. The prior approach is preferable, since a lesser volume of data must be processed. Prominent techniques include singular value decomposition (SVD), and a special case called principal component analysis (PCA).

SVD and similar techniques for feature extraction solve an optimization problem. This can be time consuming, making potential energy savings questionable. Random projections , where data is projected onto random manifolds, are a more simple and efficient dimensionality reduction technique. However, to achieve the same approximation guarantees more dimensions are needed than for SVD. Random projections preserve Euclidean distances according to the Johnson-Lindenstrauss Lemma as well as similarity computed using dot products, but random projections (as well as other dimensionality reduction techniques) do not preserve metrics such as the Manhattan distance. Therefore, some care is needed to ensure correct outcomes, when applying dimensionality reduction techniques. There is also empirical evidence comparing learning outcomes on the original data to outcomes on the data with reduced dimensionality. Unfortunately, the comparison neglects metrics relevant to energy, e.g., computation time.

Aggarwal describes dimensionality reduction with type transformation as the change of data from a more complex to a less complex type. For instance, graphs can be expressed as multidimensional data that might potentially be easier (and faster) to process. Time series can also be transformed to multi-dimensional data using the Haar Wavelet Transform or Fourier Transformation that both express the data using a (small) set of orthogonal functions. This form of data compression typically implies a loss of precision. Often, a dataset might only contain a few informative attributes and, therefore, the loss of precision might be very small, while achieving a substantial amount of data reduction. A high level understanding of the data mining task helps the data scientist choose a suitable dimensionality reduction technique. A technique might distort some instances more than others, and a small number of instances that are very different in the original context can be very similar in the space with reduced dimensions. For tasks like outlier detection this can be inacceptable, since outliers might be transformed so that they are not identifiable in the transformed data. Other tasks such as segmenting data into unspecified groups (clustering) might be less impacted by altering a few instances in a non-desirable way.

Principle #5.3: Change data representation

Data can be described in many ways without any loss of information, using lossless compression algorithms . This means that data is transformed among different representations without any effect on the minable knowledge. The green data scientist should prefer the representation that requires the least amount of storage, the least amount of computational effort to process throughout the data mining task, and the least amount of computation to create from the original data description.

A sequence of can be written more compactly as . Another form of encoding is difference encoding, where differences between two elements are stored, e.g. Difference encoding is often beneficial for time-series data, where commonly there is a strong dependency between consecutive data points. It is also possible to store only non-zero elements with indexes, e.g., the sequence 0,0,0,0,99,99 becomes 4:99, 5:99. In multiple dimensions such data structures are called sparse matrices. There are many applications where zero entries are common, e.g., document-term matrices representing textual documents and user-item matrices used to derive recommendations.

Numerous compression algorithms can be used to alter the data representation: General purpose algorithms such as Lempel-ziv as well as algorithms tailored to specific types of data. Sakr, for instance, surveys algorithms for XML data compressions. A dataset can be compressed in such a way that the entire dataset must be decompressed to access a single element. A compressed dataset might also allow for even faster access and manipulation of data than non-compressed data. For large matrices in a sparse matrix representation, for instance, some manipulations such as multiplication of two matrices are often faster. Compression and decompression also consume energy and, thus, data compression might or might not be beneficial depending on the number of required compress and decompress operations. General purpose algorithms allow to specify how much effort they should invest into finding the representation that minimizes space. Some algorithms take advantage of compressed representations and work on them directly, whereas others require an uncompressed representation. In case data is transferred across networks or is infrequently accessed, compression is even more appealing.

Principle #5.4: Accurate specification of attribute requirements

Whereas discrete attribute values stem from a fixed set of values, attributes with continuous values are

stored with a specific precision. The precision of individual attributes as well as the set of possible values can be defined by specifying an attribute type. For example, for an attribute containing temperature measurements, a data scientist might specify a precision of 0.001 degrees and a range of feasible values such as [0,100] as so called "domain constraint" in database systems . As a next step a data type can be chosen that meets these requirements and uses the least amount of storage—for instance, databases provide a set of data types according to the SQL standard, whereas programming languages usually follow the IEEE standards for floating point, integer, and other data types. The data type also determines the amount of storage and impacts the time and energy to conduct operations on data. The green data scientist should specify reasonable requirements. Choosing inappropriate types might more than double the amount of needed storage. For example, choosing an integer type (64 bits) rather than a (single) byte type (8 bits) for an array of many values leads to an increase of a factor of almost eight in memory demand.

Domain constraints depend on the data source, the range of the data, and the intended application: For sensor data, the accuracy is given by the maximum precision that seems achievable in the next years. For financial data, the needed accuracy might be given by the smallest unit, that is, one cent or one dollar. For time information, a precision up to milliseconds might not yield better outcomes than maintaining timestamps with hourly precision. For images, accuracy can be translated to the maximal resolution in terms of number of pixels or color depth that is beneficial for the analysis.

Data preparation and modeling

Principle #6: Execute common operations only once

Data preparation should be structured in such way that common preparation operations for multiple models are executed only once. For example, it can be reasonable to store a version of pre-processed data after general transformation and cleaning steps have been performed. The principle of factoring out common operations is already known, for instance, in the context of the Extract-Transform-Load (ETL) process optimization for data warehouses . The idea of storing temporary results has also been applied in the context of ETL processes and it is an integral part of the distributed data processing for Map-Reduce jobs. In both cases, the goal is fault tolerance rather than energy optimization. Strategies for identifying data processing results likely to be reused and thus worth storing have been investigated, too—for instance, for Map-Reduce jobs.

Principle #7: Choose models that enable discarding of data or low energy storage

Data lifecycle management has embraced the idea of moving data from high-cost to low-cost storage, for instance, moving data between storage tiers based on the value of data . Energy consumption and accessibility of stored data are typically negatively correlated: The easier it is to access data the more energy is required to maintain the data. Keeping data on a (magnetic) tape storage is much more energy efficient than keeping the same amount of data in the main memory of a computer. The former consumes energy only upon access, whereas main memory consumes energy even if no data is accessed. By her choices the data scientist determines the level of accessibility to data and thereby also the type of storage and amount of energy needed. The data scientist should thus be able to assess the relevance of data (over time) and assess the possibility to discard (older) data, compress (older) data, or work on summarized data. The availability of (old) data impacts the methodology that can be chosen, and the chosen methodology might also impact the data that must be stored. This is a key concern for long running systems, where data accumulates over time and models can be adjusted from time to time using newly available data. Some models can be trained incrementally using online learning algorithms, while others require the full dataset including all prior data, even in case only minor updates should be made due to new data using offline learning algorithms. For some models online as well as offline algorithms exist. Consider a system that classifies messages as spam or not spam. Such a system can be built by training a model based on previously classified messages. Since spammers adjust their strategies and style of messages, the system needs continuous updates—that is, learning. Whereas in an online learning scenario, data might be discarded after training the model, in the offline learning scenario it has to be kept.

Minimizing data access and thereby allowing to move data to energy friendly mediums is a viable option. But discarding data is a risky endeavor. What if the existing model should be replaced by a new model? Is it possible to change a model when all historic training data has been discarded? A careful assessment and management of risks is necessary. Various techniques from the domain of machine learning support reducing the need to keep data. One way is to use transfer learning by generating training data from the existing model for a new model, that is, to create labeled data in case unlabeled data is

available or can itself be generated. The disadvantage of this approach is that the generated labels are usually less accurate than the labels of the original dataset. Training data for the new model might still be highly beneficial despite transferred knowledge, but transfer learning can help reduce the amount of data needed to achieve good performance. Furthermore, training data can be enhanced by artificial training data that are a modification of existing data, thereby leading to improved results . Marginal returns decrease with additional data , and the impact in performance of having to retrain a new model might be small, even if just a small fraction of all data is retained.

Principle #8: Include only promising models and energy efficient algorithms

The traditional model selection process focuses almost exclusively on picking the model that yields the best results in terms of data mining-task-specific metrics such as accuracy or F-score for classification. A data scientist can base her model selection by comparing such metrics using empirical and theoretical comparisons (on similar datasets). The green data scientist, however, should also take into account energy consumption due to training, operating, and potentially data storage. Minor differences in task specific metrics might still be tolerable according to overall business objectives. It is not recommended to use all model and optimization algorithms as part of the computational selection process, because this leads to high energy costs. Ideally, the model candidates (and optimization algorithms) are limited to models that are likely to yield good results in terms of the desired metrics including energy efficiency. To this end, theoretical and empirical evidence should be leveraged.

A data scientist faces the choice of selecting model candidates and (hyper)parameter optimization algorithms. Energy costs are often determined by the effort to train and apply the model, that is, for predictions.

Principle #8.1: Leverage theoretical insights

Existing literature only gives limited advice on how to select the best methods for a dataset without trying them on the dataset at hand. Manning et al. advocate the use of high bias classifiers if little data is available. Properties of the learning algorithm are not the only factor impacting energy consumption. The number of hyperparameters and the effort to optimize these parameters also play a vital role. There are little theoretical foundations with respect to the best choice of hyperparameter optimization methods. The field is subject to current research . One theoretical insight is that obvious and intuitive techniques such as a systematic grid search might be inferior even to unstructured random search.

Models to describe the energy efficiency of systems and algorithms have been discussed from different perspectives such as power management , energy per low level operation (e.g., low level operations per Watt), or models involving hardware components such as CPUs and memory . However, none of these metrics seems suitable for quantifying the energy efficiency of models in the context of data mining. A data scientist usually works on a higher level of abstraction than individual hardware components and low-level CPU instructions that are the focus of many of these metrics. Theoretical computer science analyses algorithms in terms of running time. Running time, or time complexity, is the count of abstract, higher level operations needed to solve a task. The notion of time complexity can be applied to a single computer but also to a cluster of computers. In the field of parallel computing, one might simply aggregate the operations of all computers. This neglects costs due to information exchange between computers. Distributed systems such as clusters running data analytics frameworks such as Hadoop or Spark can also involve significant costs due to communication or idling (waiting for work). Generally, costs for communication, computation, and idling are tradeable . Many existing data mining algorithms are analyzed using the classical time complexity metric for a single computer, where the running time is often expressed as a function of the number of observations in a dataset and the number of dimensions. From the perspective of a green data scientist, algorithms with small time complexity seem preferable. But theoretical bounds might be coarse and, furthermore, often they neglect constants as part of the analysis process that might be of practical relevance. Therefore, empirical investigation might be more meaningful.

Principle #8.2: Leverage empirical knowledge

To the best of our knowledge a thorough comparison of learning algorithms for model parameters with respect to energy related concerns does not exist. Some works do provide empirical results for running-time of a few models, e.g., in the field of density based clustering. Running time seems to be a viable surrogate metric for measuring energy consumption of models for training and operation. For other metrics such as accuracy, multiple publications provide comparisons .

Hyperparameters often have a profound impact on model performance . To optimize hyperparameters multiple strategies exist. Some techniques try to reduce the time (and energy) for model selection by

training models on samples of data and predicting performance on the full dataset. Some optimization techniques allow to specify time constraints that guide the model selection process. Unfortunately, empirical comparisons do not report on the overall
energy consumption for training, but rather focus on other metrics such as accuracy.

5. Conclusion and future work

We introduced principles for green data mining based on the CRISP-DM methodology. Our principles apply to various phases of the process, impacting managerial decisions (e.g., make-or-buy) as well as technical questions (e.g., which model to use to conserve energy?). Creating a platform allowing to share information on model performance based on hyperparameter settings and datasets will not only be valuable for fellow data scientists, but also for
improving hyperparameter learning algorithms . Aside from empirical contributions, theoretical insights related to model selection could advance the field of green data mining. Furthermore, a detailed evaluation of the proposed principles can help in their application.

6. References

[1] Albers, S., "Energy efficient algorithms", Communications of the ACM, 53(5), 2010, pp. 86-96.
[2] Aggarwal, C. C., Data mining: The Textbook, 2015.
[3] Behrens, J. T., "Principles and procedures of exploratory data analysis", Psychological Methods, 1997.
[4] Bengio, Y., "Deep learning of representations for unsupervised and transfer learning", Proc. of ICML Workshop on Unsupervised and Transfer Learning, 2012.
[5] Bergstra, J. and Bengio, Y., "Random search for hyper- parameter optimization", Journal of Machine Learning Research, 13(Feb), 2012, pp. 281-305.
[6] Borra, S., and Di Ciaccio, A., "Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods", Computational statistics & data analysis, 54(12), 2010, pp. 2976-2989.
[7] Brooks, S., Wang, X., and Sarker, S., "Unpacking green IT: A review of the existing literature", In Proc. of the Americas Conf. on Information Systems (AMCIS), 2010.
[8] Capra, E., and Merlo, F., "Green IT: Everything starts from the software", In Proc. of European Conf. on Information Systems (ECIS), 2009.
[9] Caruana, R., and Niculescu-Mizil, A., "An empirical comparison of supervised learning algorithms", In Proc. of Int. Conf. on Machine learning, ACM, 2006.
[10] Cook, G., Pomeranz, D., Rohrbach, K., and Johnson, B., Clicking Clean: A Guide to Building the Green Internet. Greenpeace Inc., Washington, D.C, 2015.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

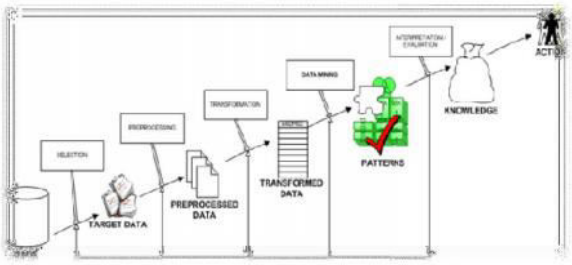
BOOSTER IN HIGH DIMENSIONAL DATA CLASSIFICATION

SANNIDHI SUBBA RAO

Dept of CSE *Malla Reddy College of Engineering* Maisammaguda, Hyderabad
sannidhisubbarao@gmail.com

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.

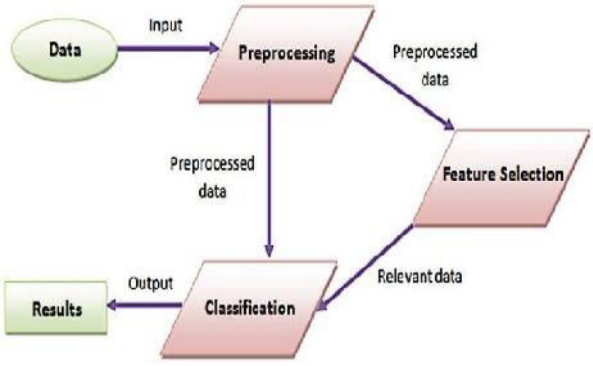
ADVANTAGES OF PROPOSED SYSTEM:

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.

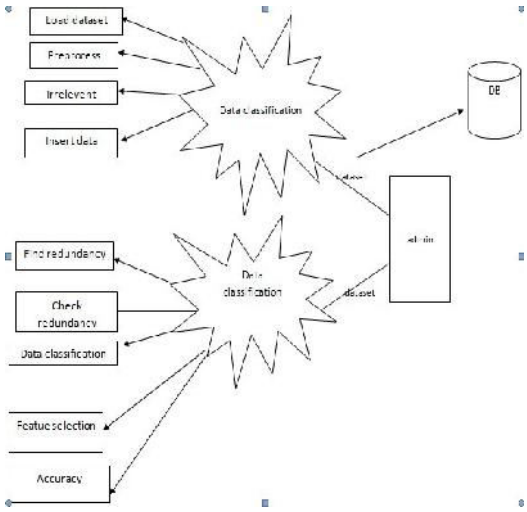
- We have noted that the classification methods applied to Booster do not have much impact on prediction accuracy and Q-statistic. Especially, the performance of mRMR-Booster was shown to be outstanding both in the improvements of prediction accuracy and Q-statistic.

SYSTEM DESIGN

SYSTEM ARCHITECTURE:



BLOCK DIAGRAM:

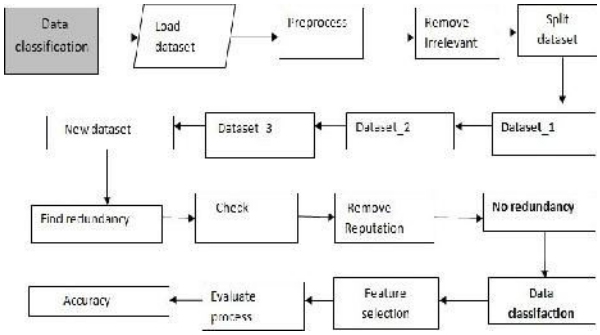


DATA FLOW DIAGRAM:

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.

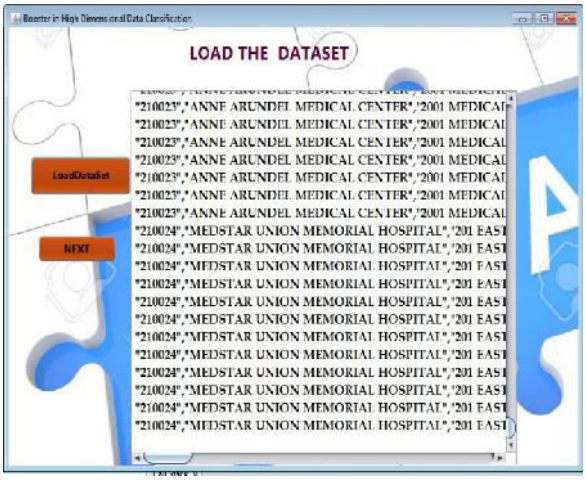
TECHNICAL FEASIBILITY

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.

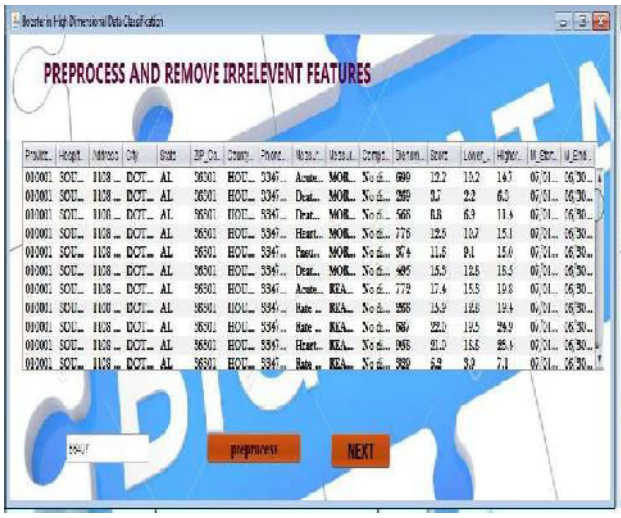
SOCIAL FEASIBILITY

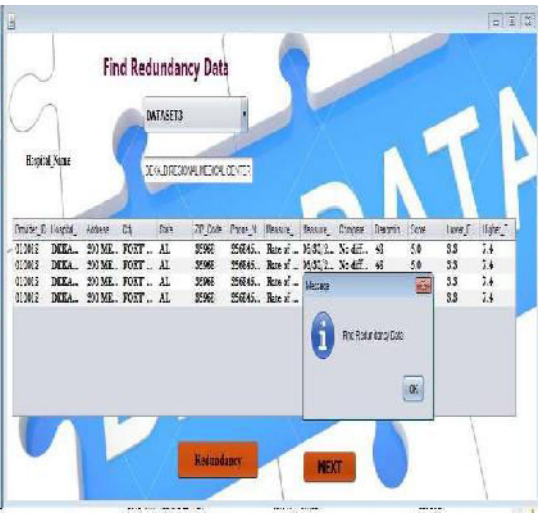
The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends 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.

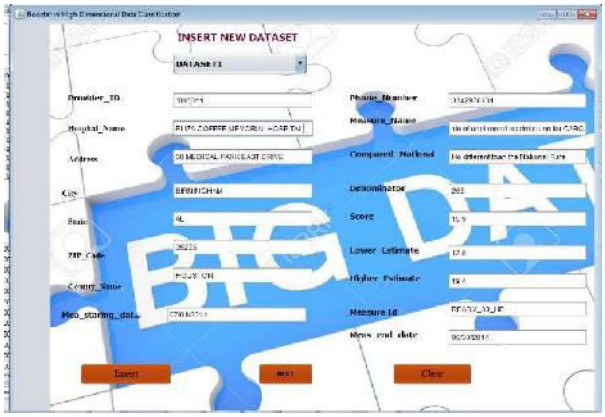
SCREEN SHOTS

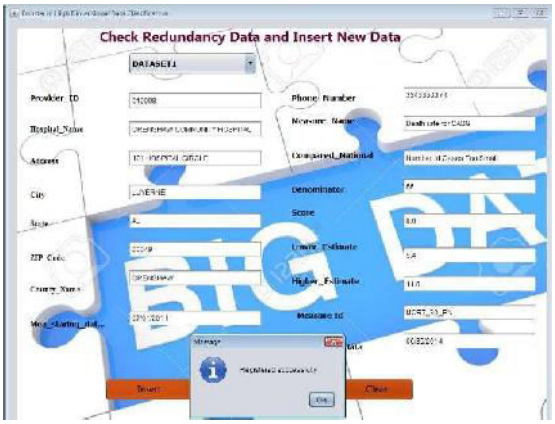


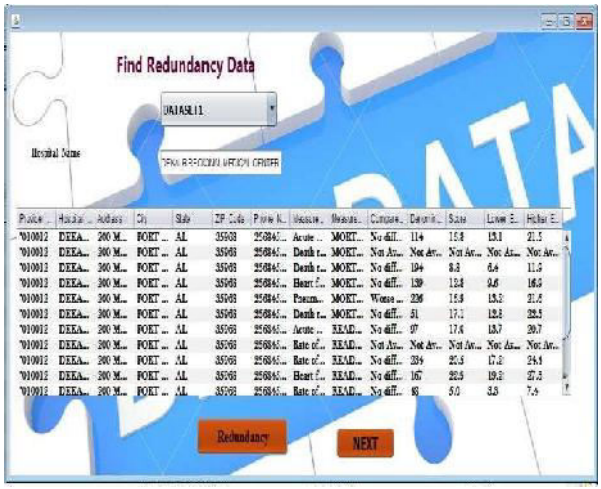
ISBN 978-93-88808-61-3

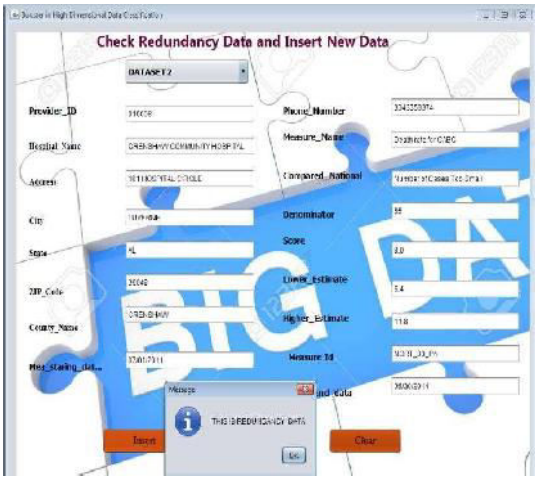




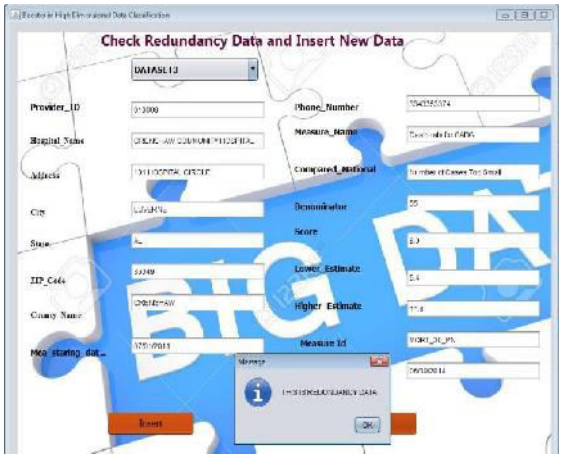


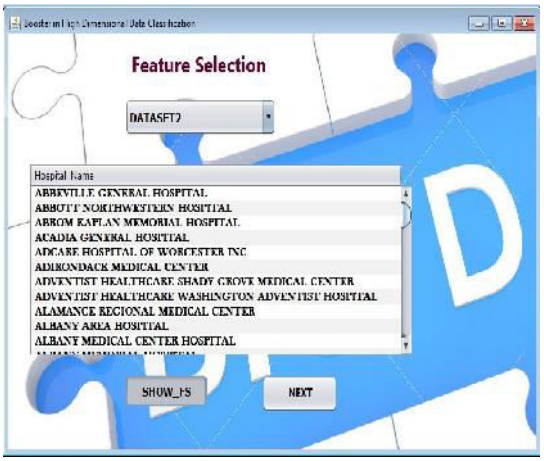


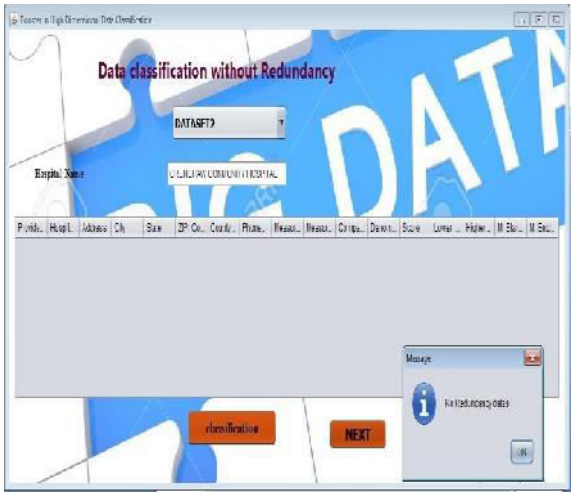


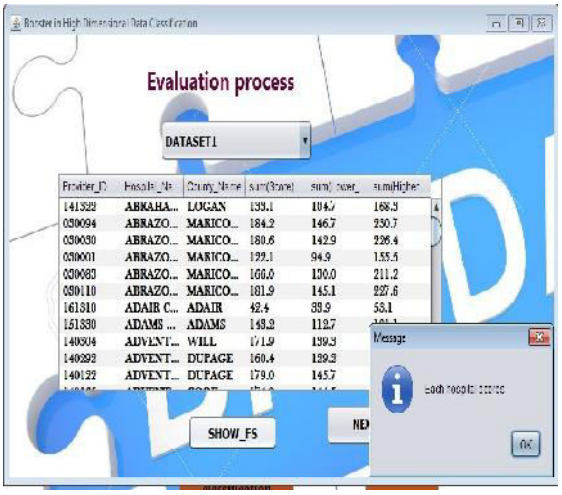


ISBN 978-93-88808-61-3

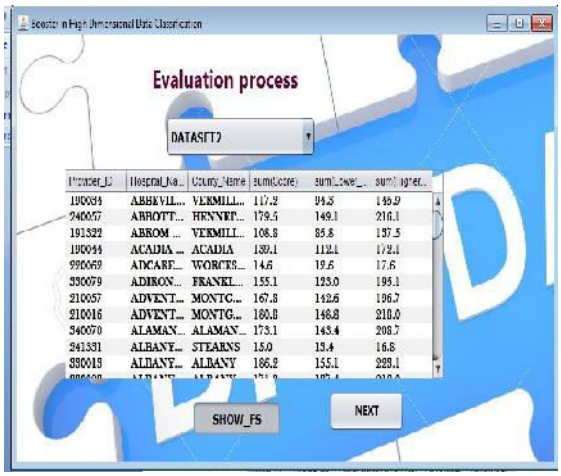


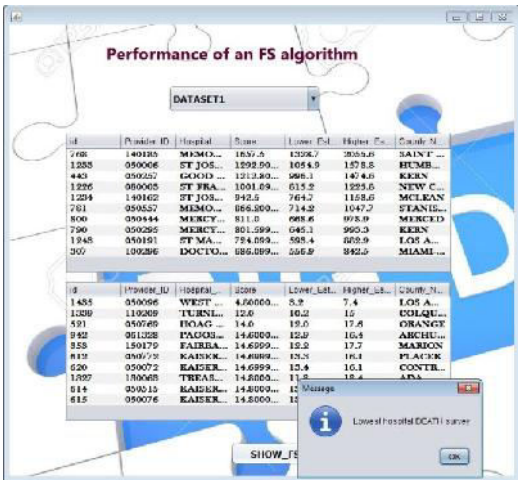


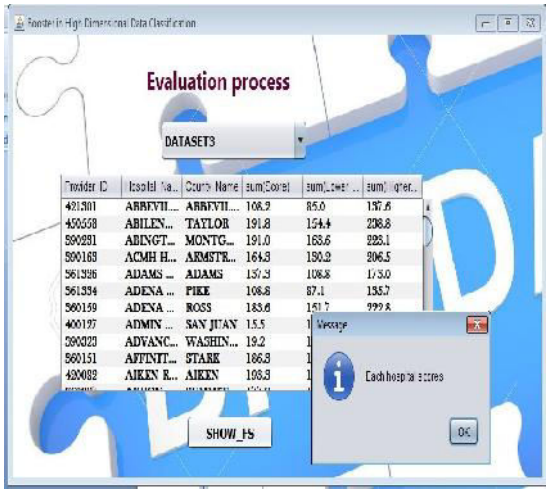


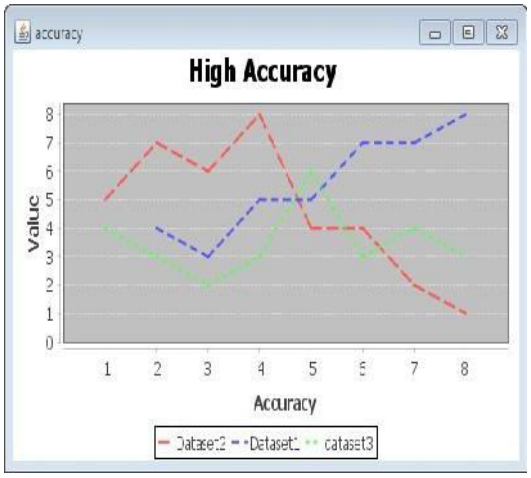


ISBN 978-93-88808-61-3

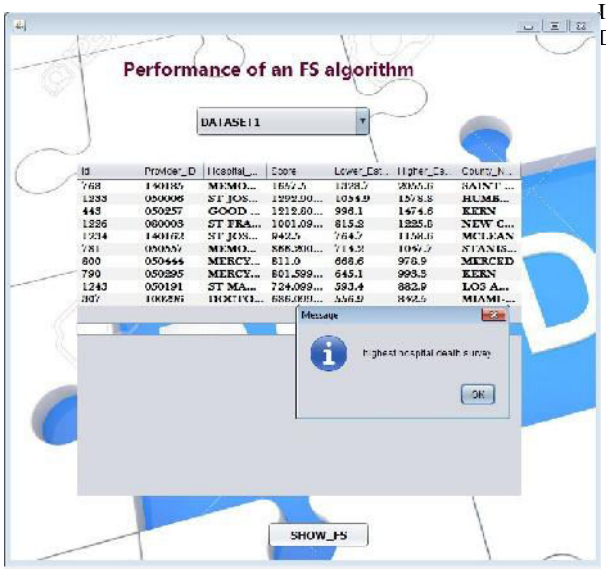








INPUT DESIGN AND OUTPUT DESIGN



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 International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Trends of Internet of Things in India

E Srinath,
Assistant Professor,
Dept. of CSE,
eslavath.sri77@gmail.com
Malla Reddy College of Engineering

Ch.Malleswar Rao
Assistant Professor
Dept. of CSE,
malleswar.538@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-winning 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 thesoil which causes attenuation of radio frequency (RF) signals. [5]

III. SYSTEM OVERVIEW

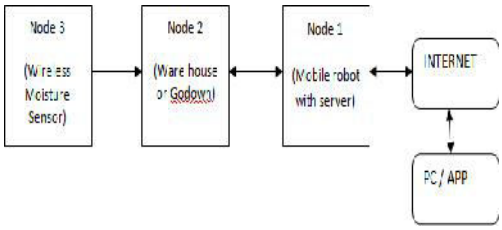


Figure 1: System overview

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.

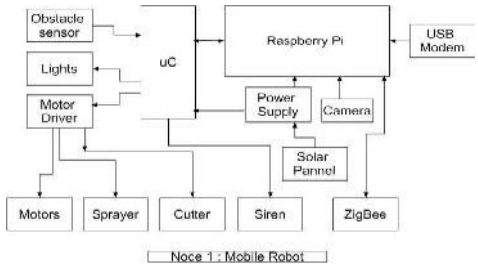


Figure 2: Node 1

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.

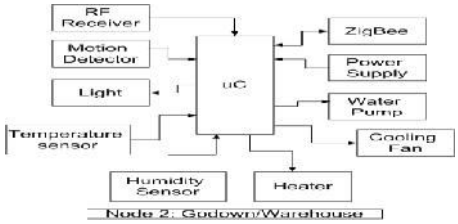


Figure 3: Node 2

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.

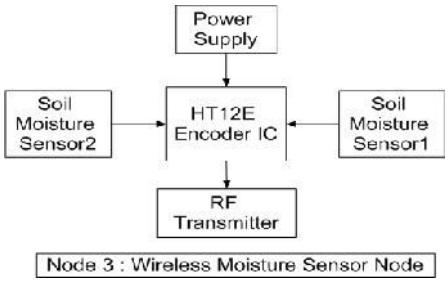


Figure 4: Node 3

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 $+150^{\circ}$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.

Software's 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). Dip Trace 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 software 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 is 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 mico controller gives signals to motor driver in order to drive the Robot.GPS module providesthe co-ordinates for the location of the robot

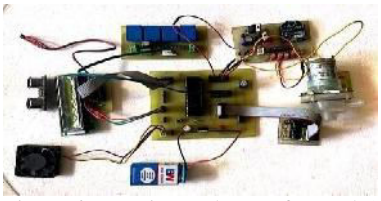


Figure 6: experimental setup for Node2

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 thecontrolaction.

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 International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Medical Internet Of Things And Bigdata In Healthcare

T.Ramchandarra,G.Naresh

ABSTRACT

A number of technologies can reduce overall costs for the prevention or management of chronic illnesses. These include devices that constantly monitor health indicators, devices that auto-administer therapies, or devices that track real-time health data when a patient self-administers a therapy. Because they have increased access to high-speed Internet and smartphones, many patients have started to use mobile applications (apps) to manage various health needs. These devices and mobile apps are now increasingly used and integrated with telemedicine and telehealth via the medical Internet of Things (mIoT). This paper reviews mIoT and big data in healthcare fields. mIoT is a critical piece of the digital transformation of healthcare, as it allows new business models to emerge and enables changes in work processes, productivity improvements, cost containment and enhanced customer experiences.

I. Introduction

The Internet of Things (IoT) is a network of physical devices and other items, embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data [1]. Its impact on medicine will be perhaps the most important, and personal, effect. By 2020, 40% of IoT-related technology will be health-related, more than any other category, making up a $117 billion market [2]. The convergence of medicine and information technologies, such as medical informatics, will transform healthcare as we know it, curbing costs, reducing inefficiencies, and saving lives.

Figure 1 illustrates how this revolution in medicine will look in a typical IoT hospital, in practice. A patient with diabetes will have an ID card that, when

scanned, links to a secure cloud which stores their electronic health record vitals and lab results, medical and prescription histories. Physicians and nurses can easily access this record on a tablet or desktop computer.

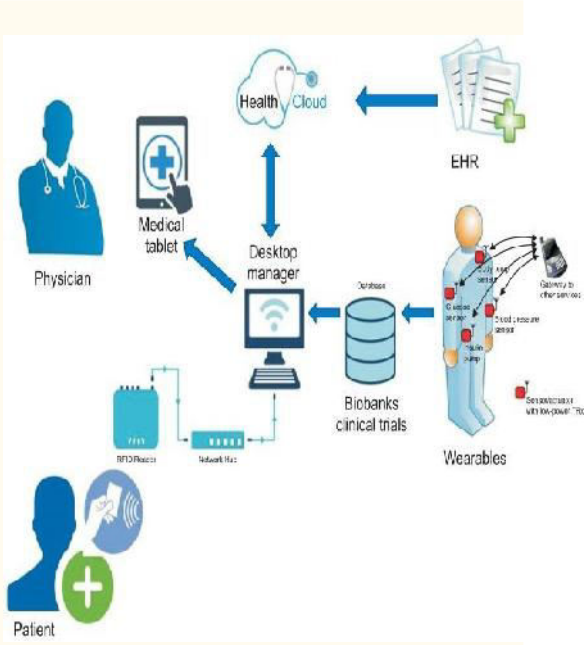


Figure 1
An illustration of how this revolution in medicine will look in a typical Internet of Things (IoT) hospital, in practice.

It sounds pretty basic, but the adoption of Electronic Health Records (EHRs) is a game changer. In less than a decade, an ink-and-paper system of managing records that goes back thousands of years will be digitized and replaced [3]. The advantages The advantages are obvious and many.,

ISBN 978-93-88808-61-3

.One of the major challenges to implementing the IoT has to do with communication; although many devices now have sensors to collect data, they often talk with the server in their own language. Manufacturers each have their own proprietary protocols, which means sensors by different makers can't necessarily speak with each other. This fragmented software environment, coupled with privacy concerns and the bureaucratic tendency to hoard all collected information, frequently maroons valuable info on data islands, undermining the whole idea of the IoT.

Precision medicine, as it's called, is a term that will be frequently heard in coming years [4]. It begins with genomics and goes through the rest of the *omics* platforms, providing multiscale data for analysis and interpretation [5]. In 2015, Intel and the Oregon Health and Science University launched a joint project, the Collaborative Cancer Cloud: a high-performance analytics platform that collects and securely stores private medical data that can be used for cancer research. Though the platform began with cancer, Intel intends to open up the federated cloud network to other institutions, including ones working on cures to diseases like Parkinson's.

Engineering simulation solutions are making medicine participatory, personalized, predictive and preventive (P4 medicine) via the medical Internet of Things (mIoT) [6].

II. IoT - The Future of Pharma?

Pharma companies long ago realized that just selling traditional medicines will not produce growth nor even sustain competitiveness. This fundamental change, known as moving 'beyond the pill', typically arises from one or two realizations: (1) medicines alone are often not enough to achieve optimal clinical outcomes for patients, and (2) as pharmaceutical pipelines dry up, 'beyond-the-pill' businesses can be valuable new sources of revenues. This has created growing interest in methods of utilizing the new technologies and business processes for development and patient care, leading to Pharma IoT.

The Pharma IoT concept involves digitalization of medical products and related care processes using smart connected medical devices and IT services (web, mobile, apps, etc.) during drug development, clinical trials and patient care. The outcomes of Pharma IoT in development and clinical trials can employ combinations of advanced technologies and services to create totally new kinds of disease treatment possibilities (e.g., Treatment 2.0).

In patient care, Pharma IoT will enable patients and healthcare professionals to use medicines with advanced sensor hardware, and craft personalized care services and processes (Product 2.0). Good examples of the Pharma IoT solutions are the connected sensor wearables for Parkinson's disease and multiple sclerosis patients, which provide medication management, improving the patient outcomes and the quality of life [7].

In addition, existing medical device products such as inhalers and insulin pens can be added to the sensor and connectivity technologies to collect data for further care analytics, and even personalized therapy [8]. All this will substantially improve personal medication and care processes, because patient care data provides new sources of innovation and competitiveness.

The transformation also involves some challenges: at the same time, pharma companies need to take into account the forthcoming European Union (EU) data protection and privacy legislation, which will give patients control of their care data

[9]. For example, patients will be allowed to transfer their care and health data across multiple service providers, leading to the emergence of totally new kinds of service platforms and business models, e.g., data brokers [10].

III. Devices and Mobile Apps for Healthcare

We are heading into the age of information, where knowledge and data will be key. We are also entering the age of the customer, in which more than ever the customer is going to determine what they want. *myTomorrows* is one example of the changing look of business models, in this case, directly connecting customers and pharma [11].

In this new age, devices and apps will be used to create a "health selfie". For example:

- *The Myo*, originally a motion controller for games, is now being used in orthopedics for patients who need to exercise after a fracture. With the aid of the Myo, patients can monitor their progress and doctors can measure the angle of movement.
- *The Zio Patch* measures heart rate and electrocardiogram (ECG) and is the US Food and Drug Administration approved [12].

Where is pharma in all this turmoil? Interestingly, there are signs that pharma is reaching out from its traditional medicine-centric approach.

- Glaxo recently announced that it is investing in electroceuticals, bioelectrical drugs that work by micro-stimulation of nerves [13].
- J&J has teamed up with Google to develop robotic surgery. In addition, they are collaborating with Philips on wearable devices such as blood pressure monitors [14].
- Novartis is working with Google (again) on sensor technologies, such as the smart lens, and a wearable device to measure blood glucose levels [15].

Sensors can provide a lot of information to support pharma development, but it is particularly important to recruit the right patients for the right clinical trials. Body sensors, once gadgets that were mainly used by athletes and runners, are now rapidly entering the general market, and consumers and pharma will soon have access to a wealth of information including not only pulse, blood pressure, ECG and respiratory rate, but also more advanced data, such as inflammation, sleep patterns, etc.

A number of mobile apps which support device handling have emerged, including myDario and SleepBot among others [16,17]. The Hacking Medicine Institute recently announced RANKED Health, a program to critically evaluate and rank health-focused applications and connected devices [18].

It has been predicted that in the near future we will look at our phone or smart watch to check health outcomes more often than we do now to check our mail or WhatsApp. A typical situation might involve an elderly person, recovering from a medical condition at home, linked to a combination of several connected services streaming data towards different parties, such as family members, tele-carer and physicians (Figure 2).

ISBN 978-93-88808-61-3

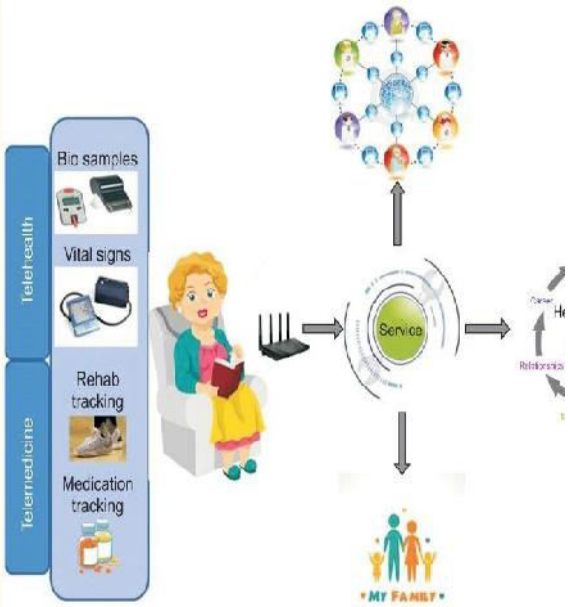


Figure 2

A typical situation involved an elderly person, recovering from a medical condition at home, linked to a combination of several connected services streaming data towards different parties, such as family members, tele-carer and physicians.

Recently it was announced that Medtronic will be partnering with a digital health app company named Canary Health to be a re-seller of its digital chronic disease management programs, including its CDC-recognized Diabetes Prevention Program, which is aimed at changing behaviors in prediabetic people. But the partnership goes beyond just reselling Canary Health's digital tools. In fact, both Canary Health and Medtronic plan to develop solutions that "leverage Medtronic's devices, services and infrastructure as well as Canary Health's suite of behavior-change programs, design expertise, and deep user engagement experience," according to a Canary Health news release [19].

One reason that Medtronic must have been attracted to Canary Health is that the company's digital tools are reimbursable. As digital health programs mature, payers are looking at innovative, yet proven, ways to reduce their cost burden for chronic diseases like diabetes.

According to the Centers for Disease Control and Prevention (CDC), people with prediabetes who take part in a structured lifestyle change program—like the one Canary Health has developed, or programs championed by Omada Health and Noom Health, among others—"can cut their risk of developing type 2 diabetes by 58% (71% for people over 60 years old)" [20]. The CDC adds that "this finding was the result of a program helping people lose 5% to 7% of their body weight through healthier eating and 150 minutes of physical activity a week" [21].

Given that diabetes is an expensive, chronic disease, hospitals, doctors, patients, and payers are equally keen to tame this epidemic. In other words, the move is helping to transform companies from simply providing care to the sick to actually delivering healthcare.

IV. Data

The driver behind all these wearable sensors is the data that is generated, and various parties are trying to bundle the data streams and obtain control. Microsoft developed the Health Vault, an e-health safe, acting as an EMR. In Holland the Radboud University Medical Center collaborated with Philips and Salesforce on *HereIsMyData*, a database where patients can store their health data and determine who can access them [22]. The role of Salesforce is interesting. The Salesforce platform powers Veeva, the customer relationship management (CRM) now widely used in pharma. This positions Salesforce to be able to bridge the gap between patient's medical data and pharma.

"Big data" is a phrase that has been used pervasively by the media and the lay public in the last several years. While

many definitions have been proposed, the common denominator seems to include the "three V's"—Volume (vast amounts of data), Variety (significant heterogeneity in the type of data available in the set), and Velocity (speed at which a data scientist or user can access and analyze the data) [23].

Defined as such, healthcare has become one of the key emerging users of big data. For example, Fitbit and Apple's ResearchKit can provide researchers access to vast stores of biometric data on users, which can then be used to test hypotheses on nutrition, fitness, disease progression, treatment success, and the like.

Most complex high dimensional data sets include imaging (photos, X-rays, MRIs, and slides), wave analysis such as EEG and ECG, audio files with associated transcripts, free text notes with natural language processing (NLP) outputs, and mappings between structured concepts such as lab tests and the Logical Observation Identifiers Names and Codes (LOINC) codes or the International Classification of Diseases-9 (ICD9) and ICD10 codes. Among the things that the data analysis should provide is the means to continuously update the annotations based on acquired knowledge, while keeping the location of the data in place.

The Centers for Medicare & Medicaid Services (CMS) have vast stores of billing data that can be mined to promote *high value care*; the same is true of private health insurers. And hospitals have attempted to reduce re-admission rates by targeting patients where predictive artificial intelligence (AI) algorithms indicate people who may be at highest risk based on an analysis of available data collected from existing patient records

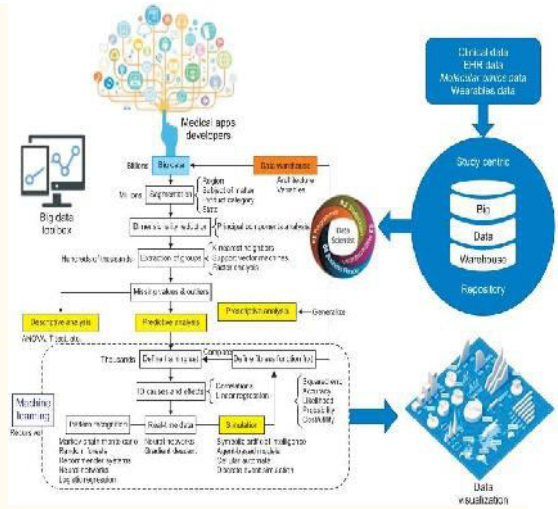


Figure 3
The Centers for Medicare & Medicaid Services (CMS) data system.

Underlying these and many other potential uses, however, are a series of technology, legal and ethical challenges relating to, among other things, privacy, discrimination, intellectual property, tort, and informed consent, as well as research and clinical ethics [24].

V. Challenges for mIoT

Leading IoT platforms must provide simple, powerful application access to IoT devices and data to help designers rapidly compose analytics applications, visualization dashboards and mIoT apps. The following are 5 key capabilities that leading platforms must enable:

(1) Simple connectivity: A good IoT platform makes it easy to connect devices and perform device management functions, scaled through cloud-based services, and to apply analytics to gain insight and achieve organizational transformation.

(2) Easy device management: A thoughtful approach to device management enables improved asset availability, increased

throughput, minimized unplanned outages and reduced maintenance costs.

(3) Information ingestion: Intelligently transform and store IoT data. APIs bridge the divide between the data and the cloud, making it easy to pull in the data that's needed. Data is ingested from diverse data sources and platforms, then the essential values are extracted using rich analytics.

(4) Informative analytics: Gain insight from huge volumes of IoT data to make better decisions and optimize operations. Apply real-time analytics to monitor current conditions and respond accordingly. Leverage cognitive analytics with both structured and unstructured data to understand situations, reason through options, and learn as conditions change. An intuitive dashboard makes it all easy to understand.

(5) Reduced risk: Act on notifications and isolate incidents generated anywhere in the company environment from a single console.

VI. Challenges for Big Data in Healthcare

The challenges fall into two main categories: fiscal/policy and technology.

Fiscal and policy issues: In a fee-for-service environment, the only way that healthcare practitioners get paid is to have face-to-face encounters with patients. This creates heavy bias against promoting technologies that streamline non-face-to-face interactions. However, as we move away from that model and more towards value-based care, where global risk-based payments are made to delivery organizations (hospitals, patient centered medical homes, accountable care organizations, etc.), then there is more incentive to use new technologies that reduce unnecessary in-office encounters. In such an environment, face-to-face encounters are actually a cost center, not a profit center, and positive health outcomes of populations are rewarded.

Technology issues: The biggest technical barrier to achieving this vision is the state of health data. Created by legacy EHR systems, health data is largely fragmented into institution-centered silos. Sometimes those silos are large, but they are still silos. Exchanging individual records between silos, using increasingly standardized vocabularies (code sets) and message formats (ADT messages, C-CDAs, even FHIR objects), is where much current effort is being directed. But that does not solve the problem of data fragmentation. More and more people in the health information exchange arena are seeing that the next generation of health technology is around aggregating data, not simply exchanging copies of individual records (the traditional query-response approach). Only by collecting the data from many different sources, normalizing that data into a consistent structure, resolving the data around unique patient identifiers as well as unique provider identifiers—only then can the data become truly useful [25].

Aggregated data has two additional advantages. (1) It solves the interoperability problem. Systems and institutions no longer need to build data bridges, and translate how the data is structured between two proprietary systems; everyone instead simply connects to a central standard API "plug." If built right, the aggregated data can be the basis for very effective AI technology. Such technology is very fast (consider Google suggestions as-you-type in a search bar, retrieving suggestions from billions of record options). (2) It is also sufficiently flexible to allow machine learning, and AI will be able to function in a real-time fashion.

ISBN 978-93-88808-61-3

VII. New Generation of Digital Health Advisors

Once a data store has been built from many different sources—EHR data, payer data, device and IoT data, patient survey responses, consumer health data—and has been integrated into a unified data structure, then AI can yield meaningful insights. AI, after all, is about pattern recognition, comparing a particular pattern of data around a given individual with similar (not necessarily identical) patterns found elsewhere, and making predictive recommendations based on what happened in those other situations. This is very much what clinicians do when exercising "clinical judgement"—identifying a pattern, taking into account medical problems, medications, labs values, personal and family history, and comparing it to similar patterns from the clinician's experience.

A new generation of "Health Coaches", Tele-Carers or Digital Health Advisors can be trained to make these AI-derived recommendations useful [26]. They need to be easy-to-use, consumer-orientated persons who can connect to the aggregated data store and the AI analytics engines that sit on top of that. They can empower consumers/patients, and reduce the demand burden on clinicians. Will they replace clinicians? No, of course not. But they will help filter the demand to those who truly need to be seen, while empowering patients with real-time, believable and personalized guidance for the more common things in day-to-day life [27].

So what stands in the way of Digital Health Advisors? Policy (how we pay for healthcare) needs to encourage self-care and facilitate healthy behaviors, rather than encourage inoffice doctor visits. And, simultaneously, health data needs to become reorganized in order to empower AI and drive the emergence of new apps and related technologies. It will be a while before we get there, but we can see the path to that new generation of healthcare technology.

Concliousn

The mIoT is revamping healthcare services, as people have started using IoT to manage their health requirements. For example, people can use IoT devices to remind them about appointments, changes in blood pressure, calories burnt and much more. One of the best parts of the IoTs in the healthcare industry is the remote health monitoring system, where patients can be monitored and advised from anywhere. Real-time location services are another major approach IoT offers. By using the service, doctors can easily track device locations, which directly reduces excess time spent. Smartphone usage is increasing rapidly, and people have started using mobile apps for almost everything. When it comes to the healthcare industry, mobile apps can improve communications between patients and doctors over a secured connection.

The primary duty of Digital Health Advisors and the clinicians will be to work collaboratively when the organization is shifting towards IoT-enabled infrastructure. Proper training and feedback are mandatory for better deployment. The traditional method of recording a patient's details, i.e., a pad of paper hanging on the patient's bed, is not going to work anymore, since such records are only accessible to a limited few, and can be lost or scrambled. This is an application where on-field mobile/tablet technology might work, since they offer hassle-free record management on the applications in the device. Health data information will be available in just a tap when information is recorded electronically, once security and privacy issues are met.

:

References

1. Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M. Internet of things for smart cities. IEEE Internet Things J. 2014;1(1):22–32. [Google Scholar]
2. Bauer H, Patel M, Veira J. The Internet of Things: sizing up the opportunity [Internet] New York (NY): McKinsey & Company; c2016. [cited at 2016 Jul 1]. Available from: http://www.mckinsey.com/industries/high-tech/our-insights/the-internet-of-things-sizing-up-the-opportunity. [Google Scholar]
3. Kruse CS, Kothman K, Anerobi K, Abanaka L. Adoption factors of the electronic health record: a systematic review. JMIR Med Inform. 2016;4(2):e19. [PMC free article] [PubMed] [Google Scholar]
4. Scheen AJ. Precision medicine: the future in diabetes care? Diabetes Res Clin Pract. 2016;117:12–21. [PubMed] [Google Scholar]
5. van Leeuwen N, Swen JJ, Guchelaar HJ, 't Hart LM. The role of pharmacogenetics in drug disposition and response of oral glucose-lowering drugs. Clin Pharmacokinet. 2013;52(10):833–854. [PubMed] [Google Scholar]
6. Flores M, Glusman G, Brogaard K, Price ND, Hood L. P4 medicine: how systems medicine will transform the healthcare sector and society. Per Med. 2013;10(6):565–576. [PMC free article] [PubMed] [Google Scholar]
7. van Uem JM, Maier KS, Hucker S, Scheck O, Hobert MA, Santos AT, et al. Twelve-week sensor assessment in Parkinson's disease: impact on quality of life. Mov Disord. 2016 May 31; doi: 10.1002/mds.26676. [Epub] [PubMed] [CrossRef] [Google Scholar]
8. Dzubur E, Li M, Kawabata K, Sun Y, McConnell R, Intille S, et al. Design of a smartphone application to monitor stress, asthma symptoms, and asthma inhaler use. Ann Allergy Asthma Immunol. 2015;114(4):341–342.e2. [PMC free article] [PubMed] [Google Scholar]
9. European Commission. Clinical trials - regulation EU No 536/2014 [Internet] Brussels: European Commission; c2016. [cited at 2016 Jul 1]. Available from: http://ec.europa.eu/health/human-use/clinical-trials/regulation/index_en.htm. [Google Scholar]
10. INNT Foundation [Internet] Amsterdam: INNT Foundation; c2016. [cited at 2016 Jul 1]. Available from: https://www.innit.foundation. [Google Scholar]
11. MyTomorrows [Internet] Amsterdam: MyTomorrows; c2016. [cited at 2016 Jul 1]. Available from: https://mytomorrows.com. [Google Scholar]
12. Tung CE, Su D, Turakhia MP, Lansberg MG. Diagnostic yield of extended cardiac patch monitoring in patients with stroke or TIA. Front Neurol. 2015;5:266. [PMC free article] [PubMed] [Google Scholar]
13. Famm K, Litt B, Tracey KJ, Boyden ES, Slaoui M. Drug discovery: a jump-start for electroceuticals. Nature. 2013;496(7444):159–161. [PMC free article] [PubMed] [Google Scholar]
14. Cuba-Gyllensten I, Gastelurrutia P, Riistama J, Aarts R, Nunez J, Lupon J, et al. A novel wearable vest for tracking pulmonary congestion in acutely decompensated heart failure. Int J Cardiol. 2014;177(1):199–201. [PubMed] [Google Scholar]
15. Senior M. Novartis signs up for Google smart lens. Nat Biotechnol. 2014;32(9):856. [PubMed] [Google Scholar]
16. MyDario.com [Internet] Burlington (MA): MyDario.com; c2016. [cited at 2016 Jul 1]. Available from: http://mydario.com/ [Google Scholar]

17. SleepBot [Internet] New York (NY): SleepBot; c2013. [cited at 2016 Jul 1]. Available from: https://mysleepbot.com/ [Google Scholar]
18. RANKED Health [Internet] place unknown: publisher unknown; [cited at 2016 Jul 1]. Available from: http://www.rankedhealth.com/about/ [Google Scholar]
19. Canary Health [Internet] Los Angeles (CA): Canary Health Inc.; c2016. [cited at 2016 Jul 1]. Available from: http://www.canaryhealth.com/ [Google Scholar]
20. Sepah SC, Jiang L, Peters AL. Long-term outcomes of a Web-based diabetes prevention program: 2-year results of a single-arm longitudinal study. J Med Internet Res. 2015;17(4):e92. [PMC free article] [PubMed] [Google Scholar]
21. Centers for Disease Control and Prevention. Lifestyle change program details [Internet] Atlanta (GA): Centers for Disease Control and Prevention; c2016. [cited at 2016 Jul 1]. Available from:]
22. Hhttp://www.cdc.gov/diabetes/prevention/lifestyle-program/experience/index.html. [Google ScholarereIsMyData [Internet] Nijmegen: Radboud University Medical Center; c2015. [cited at 2016 Jul 1]. Available from: http://www.hereismydata.com/ [Google Scholar]
23. Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, et al. Making sense of big data in health research: towards an EU action plan. Genome Med. 2016;8(1):71. [PMC free article] [PubMed] [Google Scholar]
24. Filkins BL, Kim JY, Roberts B, Armstrong W, Miller MA, Hultner ML, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res. 2016;8(3):1560–1580. [PMC free article] [PubMed] [Google Scholar]
25. Hackl WO. Intelligent re-use of nursing routine data: opportunities and challenges. Stud Health Technol Inform. 2016;225:727–728. [PubMed] [Google Scholar]
26. Foley P, Steinberg D, Levine E, Askew S, Batch BC, Puleo EM, et al. Track: a randomized controlled trial of a digital health obesity treatment intervention for medically vulnerable primary care patients. Contemp Clin Trials. 2016;48:12–20. [PMC free article] [PubMed] [Google Scholar]
27. Giraldo-Rodriguez L, Torres-Castro S, Martinez-Ramirez D, Gutierrez-Robledo LM, Perez-Cuevas R. Tele-care and tele-alarms for the elderly: preliminary experiences in Mexico. Rev Saude Publica. 2013;47(4):711–717. [PubMed] [Google Scholar]

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Prediction of Cardiac Arrest in Intensive Care Patients through Machine Learning

Rahmath Unissa and Shiva Krishna P

Abstract
Cardiac arrest is a critical health condition characterized by absence of traceable heart rate, patient's loss of consciousness as well as apnea, with inhospital mortality of *80%. Accurate estimation of patients at high risk is crucial to improve not only the survival rate, but also the quality of life as patients who survived from cardiac arrest have severe neurological effects. Existing research has focused on demonstrating static risk scores without taking account patient's physiological condition. In this study, we are implement-ing an integrated model of sequential contrast patterns using Multichannel Hidden Markov Model. These models can capture relations between exposure and control group and offer high specificity results, with an average sensitivity of 78%, and have the ability to identify patients in high risk.

Keywords
Cardiac arrest Prediction MC-HMM Sequential pattern recognition Classification

Introduction

Cardiac arrest is defined as interruption of mechanical activity of heart, which is confirmed by absence of traceable heart rate, patient's loss of consciousness, as well as apnea, according to the Utstein style. Cardiac arrest is defined as inpatient when it occurs in a hospitalized patient who had pulse on admission to the hospital [1]. Common causes of cardiac arrest are ventricular fibrillation (VF), ventricular tachycardia (VT), asystole and electrical activity of the heart without pulses. About 200,000 cases of inpatient cardiac arrest are reported each year in U.S.A. (United States of America) [2]. Cardiac arrest occurs in 1–5 per 1000 hospi-talized patients and *20% survive until their discharged [1, 3, 4]. Generally, patients at high risk of cardiac arrest have co-morbidities, which affect their health outcome and recovery after cardiac arrest [2]. Studies have shown that clinical signs of deterioration, such as hemodynamic insta-bility and respiratory distress, of patients within a period of eight hours prior to cardiac arrest could be used to avoid cardiac arrest in 84% of these [1]. However, the recognition of the causes of cardiac arrest, has been shown to increase the survival rate of patients within an hour of episode by about 29% and by 19% until their discharge [3]. Therefore, early and accurate detection of patients at-risk is critical to improve health outcome and survival rate.

Increasing use of electronic health records (EHR) leads to greater accessibility and availability of medical data. The

E. Akrivos (&) N. Maglaveras I. Chouvarda
Lab of Computing Medical Informatics and Biomedical Imaging Technologies, Medical School, Aristotle University of Thessaloniki, Thessaloniki, Greece e-mail: e_akrivos@icloud.com

E. Akrivos
Department of Internal Medicine, 424 Military Hospital of Thessaloniki, Thessaloniki, Greece

V. Papaioannou
Alexandroupolis University Hospital/Intensive Care Unit, Alexandroupoli, Greece

N. Maglaveras
McCormick School of Engineering & Applied Sciences, Department of Electrical Engineering & Computer Science, Northwestern University, Evaston, IL 60201, USA

Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II) database was developed from medical data of over 30,000 patients during 2001–2008 from Beth Israel Deaconess Medical Center in Boston. MIMIC II is the most extensive resource of intensive care unit (ICU) medical data and it is available to the public [5, 6].

Recent research used measurements of vital signs, such as blood pressure, respiratory rate, temperature and healthcare professional's opinion to model early warning scores to identify patients at high risk of cardiac arrest [7–9]. However, these researches could not predict the accurate time of cardiac arrest. DYNACARE is a model based on dynamic time series attending to predict the time of cardiac arrest [10].

The present study proposes an approach that discovers sequential contrast patterns from commonly observed measurements, such as blood pressure, respiratory rate and heart rate, transforming the classical time series data to a sequence of patterns for implementation of a classifier for cardiac arrest. Following, the classifier is used to predict the likeli-hood of a sequence of patterns to belong in cardiac arrest class. This method has been used for the prediction of Sepsis [11], but to our knowledge it is now applied to cardiac arrest prediction for the first time.

Materials and Methods

The study was conducted with data from MIMIC-II for adult patients (age 18+ on ICU admission) aged up to 90 years who were hospitalized in the Cardiological ICU and expe-rienced a recorded cardiac arrest episode according to ICD-9 (International Classification of Diseases) 427.5 for cardiac arrest. The study focused on different types of variables, such as demographic data, vitals signs, medication and laboratory measurements. Patient data was discretized in 2-hour bins. An additional requirement for each patient was to have at least 36 measurements (3 days of hospitalization) to ensure sufficient data points. There were 698 patients with a cardiac arrest diagnosis from 27,542 of MIMIC-II data-base, from which only 162 met the minimum data criteria. Patients who have been diagnosed with highrisk heart dis-eases for cardiac arrest and have not occured an event of cardiac arrest, were selected as control group, with diseases such as coronary heart disease, myocardial infarction, major heart disease, valvular heart disease, congenital heart disease and heart rhythm abnormalities such as Brugada syndrome and long QT [12–22]. The selected ICD-9 codes for these diseases was 414.01, 410.90, 429.3, 424.0, 424.1, 746.0– 746.9, 746.89 and 426.82. Similar data criteria to patients with cardiac arrest were also used in control population, with a final number of control population 5,278 patients.

Data Preparation and Preprocessing

The first step was extraction of data from Mimic Database in flat files. Quality inspection revealed a number of missing values in different data fields. Missing data was processed using the Multiple Imputation method and predictive mean matching (PMM) algorithm [23]. In order to increase similarity of considered cases, medication, demographic data and laboratory measurements were used as coefficients for PMM to predict missing values of heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, PO_2 and PCO_2.

Following, the quantization of measurements and map-ping to specific states was necessary for sequential pattern analysis, since pattern discovery methods are more effective on symbolic data types. Frequent sequence patterns methods [24–26] is used to identify patterns and frequency support in sequences between the two classes of sequence data.

Mining Sequential Contrast Patterns

Emerging patterns (EPs) are described as patterns that satisfied specific user-defined frequency rules for different classes of data. This means that in a categorized data in two categories, positive (cardiac arrest group) and negative (control group), the patterns must have a high frequency support in the positive category and a low frequency support in the negative category. Since EPs have these characteris-tics, they are considered to be distinct patterns and have the ability to distinguish the contrast between the two categories (also known as growth rate of EP). Therefore, the strength of EPs is expressed by the ratio of frequency in both classes.

Extending the above description, a sequence pattern S_p can be characterized as sequential contrast pattern if satisfy the conditions (a) and (b) depicted below in Eqs. (1) and (2)

(a) Positive support:

$$counts_{S_p}(D^{+}; g) \quad \text{a}$$

(b) Negative support:

$$counts_{S_p}(D^{-}; g) \quad \text{b}$$

where D^{+}, D^{-} two different datasets with labels, such as positive sequences and negative sequences, respectively, g is the gap-constraint, $counts_{S_p}(D; g)$ the frequency support of a

Table 1 Contrast patterns for heart rate sequences

Heart Rate patterns	Pattern-id
Tachyc < N1hr	HR1
Tachyc < Nhr	HR2
Tachyc < N1hr < N1hr	HR3
Tachyc < Tachyc	HR4

Table 2 Contrast patterns for systolic blood pressure sequences

Systolic BP patterns	Pattern-id
Nbpsys < HypotensS	SB1
HypotensS < Nbpsys	SB2
Nbpsys < Nbpsys < HypotensS	SB3

Table 3 Contrast patterns for respiratory rate sequences

Respiratory Rate patterns	Pattern-id
Bradypnoea-FP-FP	RR1
FP-FP-FP	RR2
FP-FP-Bradypnoea	RR3

sequence pattern S_p, a and b thresholds for frequency support in two datasets. Thus, discovered patterns lead to mining sequential contrast patterns, given the above characteristics, which must satisfy (a) and (b) condition [11].

In the present study, using the above description, after descretization of variables based on normal value's cut offs, resulted in contrast patterns for three variables, where a = 0.7, b = 0.5 and g = 2. Variables with contrast patterns were heart rate, systolic blood pressure and respiratory rate. Tables 1, 2 and 3 show the contrast patterns and their unique identification name with which they were replaced.

A sliding window with length equivalent to the longest pattern (length = 3) was used to transform the discrete sequences of data to sequences of contrast patterns, with purpose to use these as input data to HMM, instead of ordinary time series sequences. Table 4 shows the above transformation from a patient's sequences.

Multichannel Hidden Markov Model

Multi Channel Hidden Markov Models (MC-HMM) are an extension of the conventional form of Hidden Markov Models (HMMs) for multiple variable or channel data sequences. MC-HMM has been used on applications such as speech recognition, activity recognition, anomalous trading activities, medical events, disease interactions and fault diagnosis [27–30]. In the present study, MC-HMM was used to model interactions between multiple clinical measurements, which are represented as sequential contrast patterns. According to the theory of MC-HMM, each discrete state for each channel is individually transformed into a three-state mode, based on the markov property. Therefore, it appears that the probability of transition and emission for each state can be mapped as a mutation of the three unique states that correspond to each channel. Two MC-HMM's constructed for the two classes of data. The first MC-HMM was trained by expectation maxi-mization (EM) algorithm for patients who belong in cardiac arrest class, while the second one was trained for patients of control group.

Results

For prediction of cardiac arrest, the 8-fold cross validation method performed. For each dataset, cardiac arrest and control dataset respectively, 7 folds randomly selected used as training datasets for each model respectively and 1 fold for test set. Thus, each model was trained to find the sequences that belong to their class. Test set from cardiac arrest patient's data was containing only the sequences from observational window before the onset of cardiac arrest, for the classification purposes. Test sets from the two datasets were merged and likelihood for each patient's sequence computed for the two models. If the likelihood of the sequence patterns of the cardiac arrest patient's model was greater than the control pattern then the patient was considered to belong in class with patients at higher risk of

Table 4 Example of transformation from discrete patient sequence to sequence of contrast pattern

Discrete sequences	Variable	Contrast pattern ID
N1hr-N1hr-Tachyc-N1hr-Nhr-N1hr-….-Nhr-Tachyc-N1hr-N1hr-N1hr-Nhr-Nhr-N1hr	HR	HR1-HR3-HR2-HR3-X-X-X-X
HypotensS-Nbpsys-Nbpsys-Nbpsys-…-Nbpsys-HypotensS-Nbpsys-Nbpsys-Nbpsys-Nbpsys	SysBP	SB2-SB3-SB1-SB2-X-X-X-X
Bradypnoea-FP-RRnorm-FP-RRnorm-…-RRnorm-RRnorm-FP-RRnorm-FP-Bradypnoea-FP-RRnorm	RR	RR1-RR3-X-X-X-X-X-X

Table 5 Statistical results for 8 fold cross validation prediction

Sensitivity (mean ± SD)	Specifity (mean ± SD)
0.78 (±0.04)	0.43 (±0.02)

cardiac arrest and were categorized respectively. Table 5 shows statistical results from prediction.

Discussion

The study's results have evidence that integrating MC-HMM models with sequential contrast patterns as input data can perform well to predict cardiac arrest. A limitation in selection criteria of cardiac arrest group, which led to the reduced sample, was that patients with respiratory cardiac arrest were excluded. By discovering patterns, based on the contrast of their frequencies on data of two populations, intervention and control respectively, it is possible to interpret the difference between the two data populations. In particular, the use of a and b thresholds to calculate the growth rate of a pattern is equivalent to the odds ratio, which is used in medical research to find relationships between an exposure and an outcome. However, false positive rate is high. This issue was the result of the restricted design of population groups with ICD-9 code. In the present study, cardiac arrest diagnosis was one of the criteria for patient selection, while VT and VF diagnosis were criteria for control group. This issue poses the problem of semantically defining and selecting the correct cases within a rich database. Furthermore, the significance of a contrast pattern is determined by the growth rate, which if it is too high it creates few patterns, and if it is too low it creates patterns without significance. Therefore, creating an algorithm for optimal selection of the threshold value for pattern development is necessary in order to find important contrast patterns.

Conclusion

In this study, an attempt was made to model cardiac arrest by using an integrated framework, which was previously successfully tested to predict septic shock [11]. However, while the results are promising, it became obvious that the complexity of cardiac arrest mechanism poses many difficulties in modeling. Thus, the present study demonstrates the importance of using sequential contrast patterns to capture relations between groups.

Conflict of Interest The authors declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

References

1. Sandroni C, Nolan J, Cavallaro F, Antonelli M (2007) In-hospital cardiac arrest: incidence, prognosis and possible measures to improve survival. Intensive Care Med 33(2):237–245
2. Graham R, McCoy, MA, Schultz AM (2015) Committee on the treatment of cardiac arrest: current status and future directions, Board on Health Sciences Policy, Institute of Medicine Strategies to improve Cardiac Arrest Survival: A Time to Act. Washington (DC). National Academies Press (US), 29 Sept 2015
3. Bergum D, Haugen BO, Nordseth T, Mjølstad OC, Skogvoll E (2015) Recognizing the causes of in-hospital cardiac arrest-A survival benefit. Resuscitation. 97:91–96
4. Nolan JP, Soar J, Smith GB, Gwinnutt C, Parrott F, Power S et al (2014) Incidence and outcome of in-hospital cardiac arrest in the United Kingdom National Cardiac Arrest Audit. Resuscitation. 85 (8):987–992
5. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG et al (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101(23):E215–E220
6. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G et al (2011) Multiparameter intelligent monitoring in intensive care ii: a public-access intensive care unit database. Crit Care Med 39(5):952–960
7. Smith AF, Wood J (1998) Can some in-hospital cardiorespiratory arrests be prevented? A prospective survey. resuscitation. 37 (3):133–137
8. Hodgetts TJ, Kenward G, Vlachonikolis IG, Payne S, Castle N (2002) The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team. Resuscitation. 54(2):125–131
9. McBride J, Knight D, Piper J, Smith GB (2005) Long-term effect of introducing an early warning score on respiratory rate charting on general wards. Resuscitation. 65(1):41–44
10. Ho JC, Park Y, Carvalho CM, Ghosh J (2013) DYNACARE: dynamic cardiac arrest risk estimation. J Mach Learn Res 31:333–341

11. Ghosh S, Li J, Cao L, Ramamohanarao K (2017) Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. J Biomed Inf 66:19–31
12. Longo DL et al (2015) Cardiovascular collapse, cardiac arrest, and sudden cardiac death. In: Harrison's Principles of Internal Medicine, 19th edn, New York
13. Sudden cardiac arrest. http://www.nhlbi.nih.gov/health/health-topics/topics/scda/
14. Podrid PJ. Overview of sudden cardiac arrest and sudden cardiac death. http://www.uptodate.com/home
15. American Heart Association. Heart attack or sudden cardiac arrest: How are they different? http://www.heart.org/HEARTORG/Conditions/More/MyHeartandStrokeNews/Heart-Attack-or-Sudden-Cardiac-Arrest-How-Are-They-Different_UCM_440804_ Article.jsp-Vi55p36rTIU
16. Neumar RW, Shuster M, Callaway CW, Gent LM, Atkins DL et al (2015) Part 1: executive summary: 2015 American heart associ-ation guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation. 132(18 Suppl 2): S315-167
17. Arrhythmia. National Heart, Lung, and Blood Institute. http:// www.nhlbi.nih.gov/health/health-topics/topics/arr
18. Fuster V et al (2011) Sudden cardiac death in hurst's the heart, 13th edn. The McGraw-Hill Companies, New York
19. Goldberger AL et al (2013) Sudden cardiac arrest and sudden cardiac death in clinical electrocardiography: a Simplified Approach, 8th edn. Saunders Elsevier, Philadelphia
20. Association. AH. Ejection fraction heart failure measurement. http://www.heart.org/HEARTORG/Conditions/HeartFailure/SymptomsDiagnosisofHeartFailure/Ejection-Fraction-Heart-Failure-Measurement_UCM_306339_Article.jsp-Vi58RH6rTIU
21. Riggin EA. Allscripts EPSi. Mayo Clinic, Rochester, Minn
22. Rohren CH (expert opinion). Mayo Clinic, Rochester, Minn
23. Vink G, Frank LE, Pannekoek J, van Buuren S (2014) Predictive mean matching imputation of semicontinuous variables. Stat Neerl 68(1):61–90
24. Klema J, Novakova L, Karel F, Stepankova O (2008) Sequential data mining: A comparative case study in development of atherosclerosis risk factors. Syst Man Cybern Part C Appl Rev IEEE Trans 38(1):3–15
25. Baralis E, Bruno G, Chiusano S, Domenici VC, Mahoto NA, Petrigni C (2010) Analysis of medical pathways by means of frequent closed sequences. In: Knowledge-based and intelligent information and engineering systems, pp. 418–425
26. Berlingerio M, Bonchi F, Giannotti F, Turini F (2007) Time-annotated sequences for medical data mining. In: Seventh IEEE international conference on data mining workshops. IEEE, pp 133–138
27. Audhkhasi K, Osoba O, Kosko B (2013) Noisy hidden Markov models for speech recognition. In: The 2013 international joint conference on neural networks (IJCNN)
28. Cao L, Ou Y, Yu PS (2012) Coupled behavior analysis with applications. IEEE Trans Knowl Data Eng 24(8):1378–1392
29. Masoudi S, Montazeri N, Shamsollahi MB, Ge D, Beuche A, Pladys P, et al (2013) Early detection of apnea-bradycardia episodes in preterm infants based on coupled hidden Markov model. IEEE international symposium on signal processing and information technology
30. Zhou H, Chen J, Dong G, Wang H, Yuan H (2016) Bearing fault recognition method based on neighbourhood component analysis and coupled hidden Markov model. Mech Syst Signal Process 66– 67:568–581

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

An Efficient and Privacy-Preserving Biometric Identification Scheme in Cloud Computing

K.Ashok, M.Tech
Scholar, CSE
Department,
Malla Reddy College of Engineering,
k.ashok.kanneboina@gmail.com

ABSTRACT Biometric identification has become increasingly popular in recent years. With the develop- ment of cloud computing, database owners are motivated to outsource the large size of biometric data and identification tasks to the cloud to get rid of the expensive storage and computation costs, which however brings potential threats to users' privacy. In this paper, we propose an efficient and privacy-preserving biometric identification outsourcing scheme. Specifically, the biometric data is encrypted and outsourced to the cloud server. To execute a biometric identification, the database owner encrypts the query data and submits it to the cloud. The cloud performs identification operations over the encrypted database and returns the result to the database owner. A thorough security analysis indicates the proposed scheme is secure even if attackers can forge identification requests and collude with the cloud. Compared with previous protocols, experimental results show the proposed scheme achieves a better performance in both preparation and identification procedures.

KEYWORDS: biometric identification; data outsourcing; privacy-preserving; cloud computing

I. INTRODUCTION

BIOMETRIC identification has raised increasingly attention since it provides a promising way to identify users. Compared with traditional authentication methods based on passwords and identification cards, biometric identification is considered to be more reliable and convenient [1]. Additionally, biometric identification has been widely applied in many fields by using biometric traits such as fingerprint [2], iris [3], and facial patterns [4], which can be collected from various sensors [5]–[9].

In a biometric identification system, the database owner such as the FBI who is responsible to manage the national fingerprints database, may desire to outsource the enormous biometric data to the cloud server (e.g., Amazon) to get rid of the expensive storage and computation costs. However, to preserve the privacy of biometric data, the biometric data has to be encrypted before outsourcing. Whenever a FBI's partner (e.g., the police station) wants to authenticate an individual's identity, he turns to the FBI and generates an identification query by using the individual's biometrictraits (e.g., fingerprints, irises, voice patterns, facial patterns etc.). Then, the FBI encrypts the query and submits it to the cloud to find the close match. Thus, the challenging problem is how to design a protocol which enables efficient and privacy-preserving biometric identification in the cloud computing.

A number of privacy-preserving biometric identification solutions [10]–[17] have been proposed. However, most of them mainly concentrate on privacy preservation but ignore the efficiency, such as the schemes based on homomorphic encryption and oblivious transfer in [10], [11] for fingerprint and face image identification respectively. Suffering from performance problems of local devices, these schemes are not efficient once the size of the database is larger than 10 MB. Later, Evans et al. [12] presented a biometric identification scheme by utilizing circuit design and ciphertext packing techniques to achieve efficient identification for a

larger database of up to 1GB. Additionally, Yuan and Yu [13] proposed an efficient privacy-preserving biometric identification scheme. Specifically, they constructed three modules and designed a concrete protocol to achieve the security of fingerprint trait. To improve the efficiency, in their scheme, the database owner outsources identification matching tasks to the cloud. However, Zhu et al. [18] pointed out that Yuan and Yu's protocol can be broken by a collusion attack launched by a malicious user and cloud. Wang et al. [14] proposed the scheme CloudBI-II which used random diagonal matrices to realize biometric identification. However, their work was proven insecure in [15], [16].

In this paper, we propose an efficient and privacy-preserving biometric identification scheme which can resist the collusion attack launched by the users and the cloud. Specifically, our main contributions can be summarized as follows:

- We examine the biometric identification scheme [13] and show its insufficiencies and security weakness under the proposed level-3 attack. Specifically, we demonstrate that the attacker can recover their secret keys by colluding with the cloud, and then decrypt the biometric traits of all users.
- We present a novel efficient and privacy-preserving biometric identification scheme. The detailed security analysis shows that the proposed scheme can achieve a required level of privacy protection. Specifically, our scheme is secure under the biometric identification outsourcing model and can also resist the attack proposed by [18].
- Compared with the existing biometric identification schemes, the performance analysis shows that the proposed scheme provides a lower computational cost in both preparation and identification procedures.

The remainder of this paper is organized as follows: section II presents the models and design goals. In section III, we provide an overview and the security analysis of the previous protocol proposed by Yuan and Yu. In section IV, we present an efficient and privacy-preserving biometric identification scheme. Security analysis is presented in section V, followed by performance evaluation in section VI. In section VII, we give the related work and we show our conclusions in section VIII.

II. MODELS AND DESIGN GOALS

This section introduces the system model, attack model, design goals and the notations used in the following sections.

A. SYSTEM MODEL

As shown in Fig.1, three types of entities are involved in the system including the database owner, users and the cloud. The database owner holds a large size of biometric data (i.e., fingerprints, irises, voice, and facial patterns etc.), which is encrypted and transmitted to the cloud for storage. When a user wants to identify himself/herself, a query request is be sent to the database owner. After receiving the request, the database owner generates a ciphertext for the biometric trait and then transmits the ciphertext to the cloud for identification. The cloud server figures out the best match for the encrypted query and returns the related index to the database owner. Finally, the database owner computes the similarity between the query data and the biometric data associated with the index, and returns the query result to the user.

In our scheme, we assume that the biometric data has been processed such that its representation can be used to execute biometric match. Without loss of generality, similar to[17],[18], we target fingerprintsanduse FingerCodes [19] to represent the fingerprints. More specifically, a FingerCode consists of n elements and each element is a l-bit integer (typically $n = 640$ and $l = 8$). Given two FingerCodes $x = [x_1, x_2, .,. .x_n]$ and $y = [y_1, y_2, .,. .y_n]$, if their Euclidean distance is below a threshold s, they are usually considered as a good match, which means the two fingerprints are considered from the same person.

B. ATTACK MODEL

First of all, the cloud server is considered to be "honest but curious" as described in [13]–[15], [17]. The cloud strictly follows the designed protocol, but makes efforts to reveal privacy from both the database owner and the user. We assume that an attacker can observe all the data stored in the cloud including the encrypted biometric database, encrypted queries and matching results. Moreover, the attacker can act as a user to construct arbitrary queries.

Thus, we categorize the attack model into three levels as follows:

- Level 1: Attackers can only observe the encrypted data stored in the cloud. This follows the well-known ciphertext-only attack model [20].
- Level 2: In addition to the encrypted data stored in the cloud, attackers are able to get a set of biometric traits in the database D but do not know the corresponding ciphertexts in the database C , which is similar to the known-candidate attack model [21].
- Level 3: Besides all the abilities in level-2, attackers in level-3 can be valid users. Thus, attackers can forge as many identification queries as possible and obtain the corresponding ciphertexts. This attack follows the known-plaintext attack model [20].

A biometric identification scheme is secure if it can resist the level-α($\alpha \in \{1, 2, 3\}$) attack. Note that that if the proposed scheme can resist level-2 and level-3 attacks, it does not mean that the attacker can both be the valid user and observe some plaintexts of the biometric database simultaneously. This sophisticated attack is too strong and no effective methods is designed to defend against this kind of attack [14]. In this paper, we focus on the collusion attack between a malicious user and the cloud server. The relationship between the plaintexts of the biometric database and the ciphertexts is not known to the attacker, which is similar to the attack model proposed in [14].

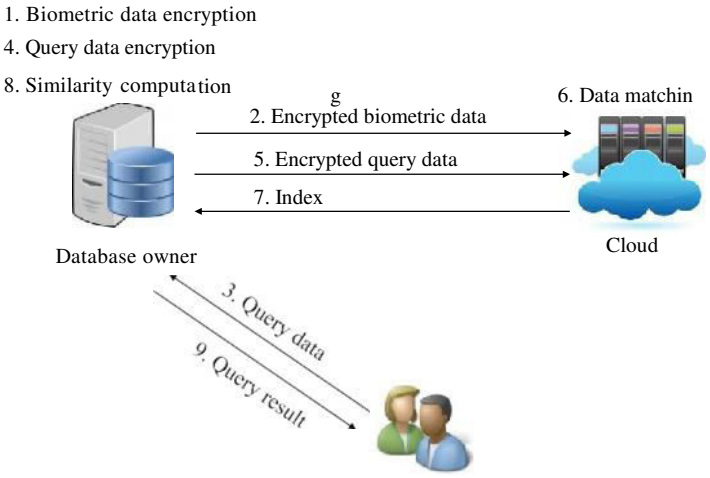


FIGURE 1. System model.

C. DESIGN GOALS

In order to achieve practicality, both security and efficiency are considered in the proposed scheme. To be more specific, design goals of the proposed scheme are described as follows:

- Efficiency: Computational costs should be as low as possible at both the database owner side and the user side. To gain high efficiency, most biometric identification operations should be executed in the cloud.
- Security: During the identification process, the privacy of biometric data should be protected. Attackers and the semi-honest cloud should learn nothing about the sensitive information.

D. NOTATIONS

Here, we list the main notations used in the remaining section as follows.

- b_i – the i-th sample FingerCode, denoted as an n-dimensional vector $b_i = [b_{i1}, b_{i2}, \cdots b_{in}]$.
- B_i – the extended sample FingerCode of b_i, denoted as an $(n+1)$-dimensional vector $B_i = [b_{i1}, b_{i2}, \cdots b_{i(n+1)}]$, where $b_{i(n+1)} = -0.5(b_{i1}^2 + b_{i2}^2 + \cdots + b_{in}^2)$.
- b_c – the query FingerCode, denoted as an n-dimensional vector $b_c = [b_{c1}, b_{c2}, \cdots b_{cn}]$.
- B_c – the extended query FingerCode of b_c, denoted as an $(n+1)$-dimensional vector $B_c = [b_{c1}, b_{c2}, \cdots b_{c(n+1)}]$, where $b_{c(n+1)} = 1$.
- W – the secret keys collection, denoted as $W = (M_1, M_2, M_3, H, R)$, where M_1, M_2 and M_3 are $(n+1) \times (n+1)$ invertible matrices, and H, R are $(n+1)$-dimensional row vectors.
- I_i – the searchable index associated with the i-th sample FingerCode b_i.
- Γ – the query FingerCodes collection constructed by the attacker, denoted as $\Gamma = (b_1, b_2, \cdots b_{t+1})$.
- B_i – the i-th extended query FingerCode constructed by the attacker, denoted as $B_i = [b_{i1}, b_{i2}, \cdots b_{i(n+1)}]$, where $b_{i(n+1)} = 1$.

III. SECURITY ANALYSIS OF YUAN AND YU'S SCHEME

In this section, we firstly describe Yuan and Yu's scheme and then give the security analysis about their scheme. To facilitate understanding of the scheme, we use $*$ to denote the elements multiplication operations, and use $\times$ to denote the matrices or vectors multiplication operations.

A. YUAN AND YU'S SCHEME

Step 1: The database owner randomly generates an $(n+1) \times (n+1)$ matrix A where $H \times A^T = 1$ and A_i is a row vector in A, $1 \le i \le (n+1)$.

Then, the database owner generates a corresponding matrix $D_i = [A^1 * b_{i1}, A^2 * b_{i2}, \cdots A^{n+1} * b_{i(n+1)}]$ to hide B_i.

After that, the database owner performs the following operations:

$$C_i = M_1 \times D_i \times M_2, \tag{1}$$

$$C_h = H \times M_1^{-1}, \tag{2}$$

$$C_r = M_3^{-1} \times R^T. \tag{3}$$

Subsequently, the database owner uploads (C_i, C_h, C_r, I_i) to the cloud, where I_i is the index of B_i.

Step 2: After Step 1 is executed, the cloud has stored many tuples in its database C. When a user requests to identify his/her identity, he/she extends b_i and then submits

the extended query B_i to the database owner. On receiving the request from the user, the database owner generates a random $(n+1)\times(n+1)$ matrix E such that $E_i \times R^T = 1$, where E_i is a row vector in matrix E and $1 \le i \le (n+1)$. The database owner then generates a corresponding matrix $F_c = [E^T_1 * b_{c1}, E^T_2 * b_{c2}, \cdots E^T_{n+1} * b_{c(n+1)}]^T$ to hide the query FingerCode B_c. The Database owner then performs the following operations:

$$C_f = M_2^{-1} \times F_c \times M_3. \tag{4}$$

Then, the database owner uploads C_f to the cloud.

Step 3: On receiving C_f, the cloud begins to search for the best match. Specifically, the cloud computes $P_i = C_h \times C_i \times C_f \times C_r$ for all encrypted biometricdatabasetocompare the Euclidean distances between b_c and b_i. Other details are eliminated since they are irrelevant for the security analysis we will describe.

B. SECURITY ANALYSIS OF YUAN AND YU'S SCHEME

In level-3 attack, an attacker has the ability to select query FingerCodes Γ of his/her interest as inputs and then tries to recover the privacy of B_i. Specifically, the attacker can compute the secret key M_2 by performing the following equation:

$$\begin{aligned} C_f \times C_r &= M_2^{-1} \times F_c \times M_3 \times M_3^{-1} \times R^T \\ &= M_2^{-1} \times F_c \times R^T \\ &= M_2^{-1} \times B_c^T. \end{aligned} \tag{5}$$

In equation 5, C_f is an $(n+1)\times(n+1)$ matrix and C_r is an $(n+1)$-dimensional vector which are both known to the attacker. B_c is an $(n+1)$-dimensional vector which can be constructed by the attacker. M_2^{-1} is one of the secret keys which is an $(n+1)\times(n+1)$ matrix but unknown to the attacker. Let S be $C \times C_f$. To recover M_2^{-1}, t query FingerCodes $\Gamma = [b_1, b_2, \cdots\cdots, b_t]$ which are extended to $[\tilde{B}^T_1, \tilde{B}^T_2, \cdots\cdots \tilde{B}^T_t]$ can be constructed, such that

$$[S_1, S_2, \cdots S_t] = M_2^{-1} \times [\tilde{B}^T_1, \tilde{B}^T_2 \cdots \tilde{B}^T_t]. \tag{6}$$

There are $(n+1)\times t$ known elements in $[S_1, S_2, \cdots\cdots, S_t]$ and $(n+1)\times t$ known elements in $[\tilde{B}^T_1, \tilde{B}^T_2, \cdots\cdots \tilde{B}^T_t]$, M_2^{-1} is a matrix with $(n+1)\times(n+1)$ unknown elements. Suppose $M_2 = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1(n+1)} \\ \vdots & \vdots & \ddots & \vdots \\ q_{(n+1)1} & q_{(n+1)2} & \cdots & q_{(n+1)(n+1)} \end{bmatrix}$, we willshowhow to recover M_2^{-1} by constructing special FingerCodes.

For the first row vector $q_1 = [q_{11}, q_{12}, \cdots, q_{1(n+1)}]$ in M_2^{-1}, the adversary constructs two special vectors as $\tilde{B}^T_1 = [1, 0, \cdots, -0.5]$, and $\tilde{B}^T_2 = [2, 0, \cdots, -2]$. Then, the attacker can compute as follows:

$$\begin{aligned} 1 * q_{11} - 0.5 * q_{1(n+1)} &= S_{11}, \\ 2 * q_{11} - 2 * q_{1(n+1)} &= S_{21}. \end{aligned} \tag{7}$$

From equation 7, it is easy to compute q_{11} and $q_{1(n+1)}$. Following the same analysis, the attacker can obtain all the elements in M_2^{-1} by constructing other special vectors.

After recovering M_2^{-1}, the attacker can compute the biometric data as follows:

$$\begin{aligned} C_h \times C_i &= H \times M^{-1} \times M \times D_i \times M_2 \\ &= H \times D_i \times M_2 \\ &= B_i \times M_2. \end{aligned} \tag{8}$$

In equation 8, C_h and C_i are known by the attacker. M_2 is the secret key which is recovered by the above foregoing. Therefore, the attacker can recover B_i.

IV. A NOVEL BIOMETRIC IDENTIFICATION SCHEME

In this section, we show the details of the proposed biometric identification scheme.

A. OVERVIEW

We construct a novel biometric identification scheme to address the weakness of Yuan and Yu's scheme [13]. To achieve a higher level of privacy protection, a new retrieval way is constructed to resist the level-3 attack. Moreover, we also reconstruct the ciphertext to reduce the amount of uploaded data and improve the efficiency both in the preparation and identification procedures.

In the remaining part of this section, we will introduce the preparation process and the identification process.

B. PREPARATION PROCESS

In the preparation process, b_i isthe i-th sample feature vector derived from the fingerprint image using a feature extraction algorithm [19]. To be more specific, b_i is an n-dimensional vector with l bits of each element where $n = 640$ and $l = 8$.

For ease of identification, b_i is extended by adding an $(n+1)$-th element as B_i. Then, the database owner encrypts B_i with the secret key M_1 as follows:

$$C_i = B_i \times M_1. \tag{9}$$

The database owner further performs the following operation:

$$C = M^{-1} \times H. \tag{10}$$

Each FingerCode B_i is associated with an index I_i. After execute the encryption operations, the database owner uploads (C_i, C_h, I_i) to the cloud.

C. IDENTIFICATION PROCESS

The identification process includes the following steps:

Step 1: When a user has a query fingerprint to be identified, he/she first gets the query FingerCode b_c derived from the query fingerprint image. The FingerCode b_c is also an n-dimensional vector. Then, the user sends b_c to the database owner.

Step 2: After receiving b_c, the database owner extends b_c to B_c by adding an $(n+1)$-th element equals to 1. Then the database owner randomly generates an $(n+1) \times (n+1)$ matrix E. The i-th row vector $E_i = [E_{i1}, E_{i2}, \cdots E_{i(n+1)}]$ is set as a random vector, where the $(n+1)$-th element is $(1 - \sum_{j=1}^{n} E_{ij} * H_j)/H_{n+1}$, $1 \leq i \leq (n+1)$. After that, the database owner performs the following computation to hide B_c:

$$F_c = [E^T * b_{c1}, E^T * b_{c2}, \cdots E^T * b_{c(n+1)}]_T. \quad (11)$$

To securely send F_c to the cloud, the database owner needs to encrypt F_c with the secret keys and a random integer $r (r > 0)$. The computation is performed as follows:

$$C_f = M_1^{-1} \times r \times F_c \times M_2. \quad (12)$$

Then, the database owner sends C_f to the cloud for identification.

Step 3: After receiving C_f from the database owner, the cloud begins to search the FingerCode which has the minimum Euclidean distance with the query FingerCode B_c. P_i denotes the relative distance between B_i and B_c as follows:

$$\begin{aligned} P_i &= C_i \times C_f \times C_h \\ &= B_i \times M_1 \times M_1^{-1} \times r \times F_c \times M_2 \times M_2^{-1} \times H^T \\ &= B_i \times r \times F_c \times H^T \\ &= \sum_{j=1}^{n+1} r * b_{ij} * b_{cj}. \end{aligned} \quad (13)$$

In equation 13, the computation result is an integer, which can be used to compare two FingerCodes. For example, to compare the query b_c with two FingerCodes, say b_i and b_z, the cloud computes P_i and P_z, and performs the following operation, where $1 \leq i, z \leq t$, $i \neq z$:

$$\begin{aligned} P_i - P_z &= \sum_{j=1}^{n+1} r * b_{ij} * b_{cj} - \sum_{j=1}^{n+1} r * b_{zj} * b_{cj} \\ &= \left(\sum_{j=1}^{n} r * b_{ij} * b_{cj} - 0.5 \sum_{j=1}^{n} r * b_{ij}^2 \right) \\ &\quad - \left(\sum_{j=1}^{n} r * b_{zj} * b_{cj} - 0.5 \sum_{j=1}^{n} r * b_{zj}^2 \right) \\ &= 0.5r(dist_{ic}^2 - dist_{zc}^2). \end{aligned} \quad (14)$$

As shown in equation 14, if $P_i - P_z > 0$, the cloud learns that b_i matches the query FingerCode much better than b_z. After repeating the operations for the encrypted FingerCode database C in the cloud, the ciphertext C_i which has the minimum Euclidean distance with b_c can be found. The cloud further gets the corresponding index I_i according to the tuple (C_i, C_h, I_i) and sends it back to the database owner.

Step 4: After receiving the index I_i, the database owner gets the corresponding sample FingerCode b_i in the database D and calculates the accurate Euclidean distance between b_i and b_c as $dist_{ic} = \sum_{j=1}^{n} (b_{ij} - b_{cj})^2$. Then, the database owner compares the Euclidean distance with the standard threshold. If the distance is less than the threshold value, the query is identified. Otherwise, the identification fails.

Step 5: Finally, the database owner returns the identification result to the user.

V. SECURITY ANALYSIS

In this part, we first prove that our scheme is secure under level-2 and level-3 attacks, and then we will show the proposed scheme can resist the attack proposed by Zhu et al [18].

A. SECURITY ANALYSIS UNDER LEVEL-2 ATTACK

According to the attack scenario 2, an attacker can obtain some plaintexts of the biometric database, but does not know the corresponding ciphertexts.

We consider C_i which is obtained by multiplying B_i and M_1. Since the mapping relationship between B_i and C_i is not known, it is impossible for the attacker to compute B_i and M_1.

B. SECURITY ANALYSIS UNDER LEVEL-3 ATTACK

In the level-3 attack, besides the knowledge of encrypted data in the cloud, the attacker can forge a large number of query FingerCodes Γ as inputs. In the following, we will show the proposed scheme is secure by proving that the secret keys cannot be recovered.

When colluding with the cloud, the attacker gets C_f and C_h, and then performs the following operation:

$$\begin{aligned} C_f \times C_h &= M_1^{-1} \times r \times F_c \times M_2 \times M_2^{-1} \times H^T \\ &= M_1^{-1} \times r \times F_c \times H^T \\ &= M_1^{-1} \times r \times B_i^T. \end{aligned} \quad (15)$$

In equation 15, since r is a positive random integer in identification process, the attacker cannot compute the secret key M_1^{-1} directly.

Pretending a valid user, the attacker can construct t query FingerCodes $\Gamma = [\tilde{b}_1, \tilde{b}_2, \cdots \tilde{b}_t]$ extended as $[\tilde{B}_1, \tilde{B}_2, \cdots \tilde{B}_t]$ for identification, which introduces a set of positive random values r_j and C_{fj}, $1 \leq j \leq t$. Let $\tilde{P}_j$ be the value of $C_{fj} \times C_h$. The attacker computes $\tilde{P}_j$ as follows:

$$\tilde{P}_j = M_1^{-1} \times r_j \times \tilde{B}_j^T. \quad (16)$$

After constructing t equations, we have:

$$\begin{aligned} \tilde{P} &= M_1^{-1} \times [\tilde{B}_1^T, \tilde{B}_2^T, \cdots \tilde{B}_t^T] \times \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_t \end{bmatrix} \\ &= M_1^{-1} \times \tilde{B} \times R. \end{aligned} \quad (17)$$

Here $[\tilde{B}_1^T, \tilde{B}_2^T, \cdots \tilde{B}_t^T]$ is denoted as B, $\begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_t \end{bmatrix}$ is denoted as R. In this equation, P is an $(n+1) \times t$ matrix known to the attacker, $\tilde{B}$ is an $(n+1) \times t$ matrix constructed by the attacker, R is an $t \times t$ matrix, since r_j is a random positive integer, it is unknown to the attacker.

We then demonstrate that the attacker cannot recover M_1 according to **Theorem 1.**

Theorem 1. Assume after t equationsare constructed, M_1 cannot be computed in $\tilde{P} = M_1^{-1} \times \tilde{B} \times R$. When $(t+1)$ equationsare constructed, the following equation holds, and M_1 cannot be recovered.

$$\tilde{P}|\tilde{P}_{t+1} = M_1^{-1} \times [\tilde{B}|\tilde{B}^T_{t+1}] \times \begin{bmatrix} R & 0 \\ 0 & r_{t+1} \end{bmatrix}. \quad (18)$$

Proof. This theorem is proven with *the inductive method.* When $t = 1$, M_1 cannot be computed in equation 16. Assume the equation 17 holds, where $(t > 1)$. When $(t+1)$ query FingerCodes are constructed, we obtain:

$$\tilde{P}, \tilde{P}_{t+1} = [M_1^{-1} \times \tilde{B}, M_1^{-1} \times \tilde{B}^T_{t+1}] \times \begin{bmatrix} R & 0 \\ 0 & r_{t+1} \end{bmatrix}. \quad (19)$$

For $(t+1)$-th query FingerCode $\tilde{B}_{t+1}$, we have

$$\tilde{P}_{t+1} = M_1^{-1} \times \tilde{B}^T_{t+1} \times r_{t+1}. \quad (20)$$

From equation 20, we have

$$\tilde{B}_{t+1} \times (M_1^{-1})^T = (r_t^{-1})^T \times \tilde{P}^T_{t+1}. \quad (21)$$

Let $(d_1, d_2, \cdots d_{n+1})$ be the vector $(r^{-1})^T \times \tilde{P}^T_{t+1}$ where $d_j = (r_t^{-1})^T \times P_{(t+1)j}$. Let $(M_1^{-1})^T = (m_1^T, m_2^T, \cdots, m_{n+1}^T)$, where m_j denotes a row vector in M_1^{-1}, $1 \le j \le (n+1)$. The following equations hold:

$$\tilde{B}_{t+1} \times (m_1^T, m_2^T, \cdots m_{n+1}^T)^T = (d_1, d_2, \cdots, d_{n+1}), \quad (22)$$

$$\tilde{B}^T_j \times m_j = d_j \quad (23)$$

Equation 23 is a typical non-linear homogeneous equation. Since the rank of $\tilde{B}_{t+1}$ is $r(\tilde{B}_{t+1})$, we assume the result is $\alpha_1\beta_1 + \alpha_2\beta_2 + \cdots + \alpha_{(n+1-r(\tilde{B}_{t+1}))}\beta_{n+1-r(\tilde{B}_{t+1})}$. We further state the special solution of equation 23 is β^* which satisfies the formula $\tilde{B}_{t+1} \times m_j = d_j$. Because $d_j = (r_{t+1})_j \times \tilde{P}^T_{t+1} \times (r^{-1})^T$, β^* is included in the special solution of m_j^T. For m_j^T in matrix $(M_1^{-1})^T$, the particular solution of m_j^T is $\alpha_1\beta_1 + \alpha_2\beta_2 + \cdots + \alpha_{(n+1-r(\tilde{B}_{t+1}))}\beta_{n+1-r(\tilde{B}_{t+1})} + \beta^*$. Since r is a random integer, the special solution β^* is uncertain as well, which means the attacker cannot derive the exact particular solution for m_j^T in $(M_1^{-1})^T$. □

Therefore, when $(t+1)$ query FingerCodes are constructed, the secret key M_1 cannot be computed by the attacker as well.

As discussed above, the attacker cannot recover the secret key even if he is a malicious user. Therefore, the attacker cannot recover the biometric data as well.

Moreover, we compare our scheme with the schemes proposed in [13] and [14]. According to Table 1, other schemes have some weaknesses, while our scheme is secure under all the three level attacks

C. SECURITY ANALYSIS UNDER THE ATTACK PROPOSED BY ZHU ET AL.

Zhu et al. [18] showed an attack for Yuan and Yu's scheme. In their attack, the attacker observes the cloud and gets the values of relative distance. According to the equation 1, 2, 3, 4, the relative distance in Yuan and Yu's scheme can be computed as follows:

$$\begin{aligned} P_i &= C_h \times C_i \times C_f \times C_r \\ &= H \times M_1^{-1} \times M_1 \times D_i \times M_2 \\ &\times M_3^{-1} \times F_c \times M_3 \times M_2^{-1} \times R^T \\ &= H \times D_i \times F_c \times R^T \\ &= \sum_{j=1}^{n+1} b_{ij} * b_{cj} \\ &= B_i \times B_c^T \end{aligned} \quad (24)$$

As shown in equation 24, P_i is an integer which the attacker can get in the cloud, B_c is the extended query FingerCode which can be constructed by the attacker pretending to be a user. B_i is the extended sample FingerCode which is sensitive and should not be leaked. To recover B_i, the attacker can construct t query FingerCodes $\Gamma = [\tilde{b}_1, \tilde{b}_2, \ldots \tilde{b}_t]$ extended as $[B_1, B_2, \cdots B_t]$ for identification. P_{ij} denotes the relative distance between the sample FingerCode B_i and the query FingerCode $\tilde{B}_j$ where $1 \le j \le t$. Then, the attacker has:

$$[\tilde{P}_{i1}, \tilde{P}_{i2}, \cdots \tilde{P}_{it}] = [b_{i1}, b_{i2}, \cdots b_{i(n+1)}] \times [\tilde{B}^T_1 \tilde{B}^T_2 \cdots \tilde{B}^T_t]. \quad (25)$$

In this equation, $\tilde{P}_{ij}$ and $\tilde{B}_j$ are known to the attacker. For each element in B_i, it can be recovered if t equations are built, where $t > (n+1)$.

Then, we demonstrate the proposed scheme is secure under the attack proposed by Zhu et al. In the proposed scheme, $\tilde{P}_{ij}$ is set as the relative distance between B_i and $\tilde{B}_j$.

$$\begin{aligned} \tilde{P}_{ij} &= C_i \times C_{fj} \times C_{h_j} \\ &= r_j \times B_i \times \tilde{B}^T. \end{aligned} \quad (26)$$

r_j is the j-th positive random integer in t identification processes. The attacker constructs t query FingerCodes and gets the equation as follows:

TABLE 1. Security comparison with other schemes.

Schemes	Level 1 attack	Level 2 attack	Level 3 attack
Yuan and Yu's scheme [13]	Yes	Yes	No
Wang et al.'s scheme [14]	Yes	Yes	No
Our scheme	Yes	Yes	Yes

$$[\tilde{P}_{i1}, \tilde{P}_{i2}, \cdots \tilde{P}_{it}] = [b_{i1}, b_{i2}, \cdots b_{i(n+1)}] \times [r_1\tilde{B}_1^T, r_2\tilde{B}_2^T, \cdots r_t\tilde{B}_t^T]$$

$$= [b_{i1}, b_{i2}, \cdots b_{i(n+1)}] \times [B_1, B_2, \cdots B_t]^T \times \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_t \end{bmatrix} \quad (27)$$

In this equation, r_j is a positive random integer which is unknown to the attacker. For every element in B_i, after t computations, the attacker can only get the value of $r_j * b_{iq}$ where $t > (n+1)$, $1 \le q \le (n+1)$. For the reason that r_j is a random integer, $r_j * b_{iq}$ is also unexpected which means the attacker cannot acquire B_i. Thus, the proposed scheme can resist the attack proposed by Zhu et al.

VI. PERFORMANCE ANALYSIS

To evaluate the performance of the proposed scheme, we implement a cloud-based privacy-preserving fingerprint identification system. For the cloud, we use 2 nodes with 6-core 2.10 GHz Intel Xeous CPU and 32GB memory. We utilize a laptop with an Intel Core 2.40 GHz CPU and 8G. Similar to [13] and [14], the query FingerCodes are randomly selected from the database which is constructed with random 640-entry vectors.

A. COMPLEXITY ANALYSIS

Table 2 summarizes the computation and communication costs on the data owner side, cloud server and users in our scheme and the schemes in [13] and [14]. In this work, each matrix multiplication costs $O(n^3)$, where n denotes the dimension of a FingerCode, and the sorting cost of fuzzy Euclidean distances has time complexity of $O(m \log m)$. As illustrated in Table 2, our scheme has lower complexities in the preparation phase. That is, more computation and bandwidth costs can be saved for the database owner. In the identification phase, the computation complexity of our scheme is lower than that in [14]. The reason is that our scheme performs vector-matrix multiplication operations to find the close match, while [14] needs to execute matrix-matrix multiplication operations. Although the complexity of our scheme is the same as that in [13], we emphasize that [13] sacrifices the substantial security to achieve such fast computation of P_i. Moreover, our scheme executes fewer multiplication operations, and thus obtains better performance.

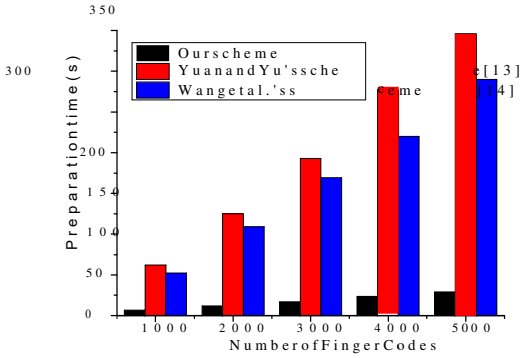


FIGURE 2. Time costs in the preparation phase.

B. EXPERIMENTAL EVALUATION

Preparation phase. Fig. 2 and Fig. 3 show the computation and communication costs in the preparation phase with the number of FingerCodes varying from 1000 to 5000. As shown in Fig.2, in our scheme, registering 5000 FingerCodes needs 29.37s, which can save about 88.85% and 90.58% time cost compared with [13] and [14] respectively. The reason is when encrypting a sample FingerCode, in our scheme, only one matrix is needed which leads to fewer matrix multiplication operations. Fig. 3 shows the bandwidth costs of the three schemes. Since the data outsourced to the cloud is in the form of vectors in comparison with matrices in the other two schemes, the communication cost in our scheme is much less than [13], [14].

Identification phase. Fig.4 and Fig. 5 show the computation and communication costs in the identification phase with the number of FingerCodes ranges from 1000 to 5000. As demonstrated in Fig. 4, all schemes grow linearly as the size of database increases. As in our scheme fewer matrix multiplication operations are used than [13], it can save about 56% time cost. Compared with [14], the identification time can be saved as much as 84.75%, since the vector-matrix multiplication rather than the matrix-matrix multiplication operation is executed. The bandwidth costs of the three schemes, as shown in Fig. 5, are almost the same. The reason is that all schemes need to transmit a matrix in the identification phase.

TABLE 2. A summary of complexity costs. In the table, m denotes the number of FingerCodes in the biometric database; n m.

		Phases	Yuan and Yu's scheme [13]	Wang et al.'s scheme [14]	Our scheme
Computation	Database owner	Preparation	$O(mn^3)$	$O(mn^3)$	$O(mn^2)$
		Identification	$O(n^3)$	$O(n^3)$	$O(n^3)$
		Retrieval	$O(n)$	$O(n)$	$O(n)$
	Cloud server	Identification	$O(mn^2 + m \log m)$	$O(mn^3 + m \log m)$	$O(mn^2 + m \log m)$
	User	Identification	/	/	/
Communication	Database owner	Preparation	$O(mn^2)$	$O(mn^2)$	$O(mn)$
		Identification	$O(n^2)$	$O(n^2)$	$O(n^2)$
		Retrieval	$O(1))$	$O(1)$	$O(1)$
	Cloud server	Identification	/	/	/
		Retrieval	$O(1)$	$O(1)$	$O(1)$
	User	Identification	$O(1)$	$O(1)$	$O(1)$

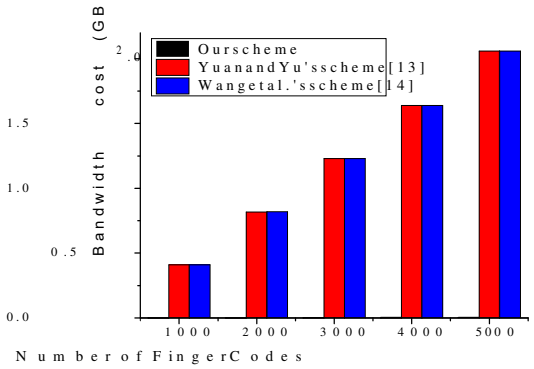


FIGURE 3. Bandwidth costs in the preparation phase.

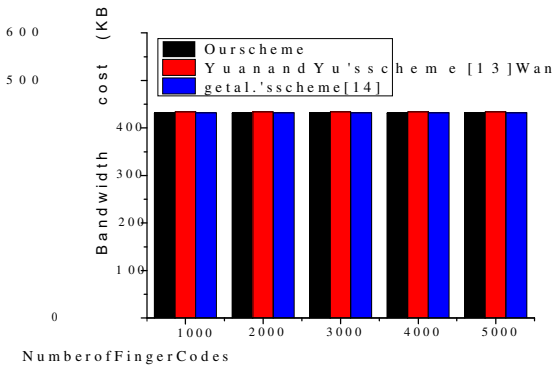


FIGURE 5. Bandwidth costs in the identification phase.

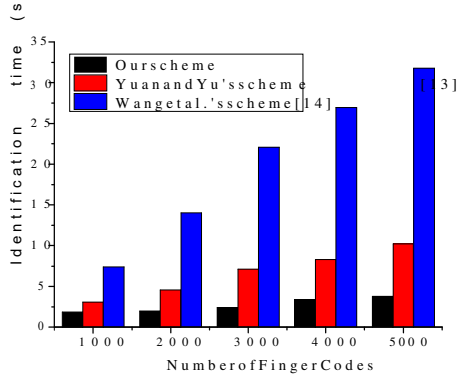


FIGURE 4. Time costs in the identification phase.

VII. RELATED WORKS

Related works on privacy-preserving biometric identification are provided in this section. Recently, some efficient biometric identification schemes have been proposed. Wang and Hatzinakos proposed a privacy-preserving face recognition scheme [22]. Specifically, a face recognition method is designed by measuring the similarity between sorted index numbers vectors. Wong and Kim [23] proposed a privacy-preserving biometric matching protocol for iris codes verification. In their protocol, it is computationally infeasible for a malicious user to impersonate as an honest user. Barni et al. [10] presented a FingerCode identification protocol based on the Homomorphic Encryption technique. However, all distances are computed between the query and sample Fingercodes in the database, which introduces too much burden as the size of fingerprints increases. To improve the efficiency, Evans et al. [12] proposed a novel protocol which reduces the identification time. They used an improved Homomorphic encryption algorithm to compute the Euclidean distance and designed novel garbled circuits to find the minimum distance. By exploiting a backtracking protocol, the best match FingerCode can be found. However, in [12], the whole encrypted database has to be transmitted to the user from the database server. Wong et al. [24] proposed an identification scheme

based on kNN to achieve secure search in the encrypted database. However, their scheme assumes that there is no collusion between the client side and cloud server side. Yuan and Yu [13] proposed an efficient privacy-preserving biomet- ric identification scheme. However, Zhu et al. [18] pointed out their protocol can be broken if a malicious user colludes with the cloud server in the identification process. Based on [13], Wang et al. presented a privacy-preserving biometric identification scheme in [14] which introduced random di- agonal matrices, named CloudBI-II. However, their scheme has been proven insecure in [15], [16]. Recently, Zhang et al. [17] proposed an efficient privacy-preserving biometric identification scheme by using perturbed terms.

VIII. CONCLUSION

In this paper, we proposed a novel privacy-preserving biometric identification scheme in the cloud computing. To realize the efficiency and secure requirements, we have designed a new encryption algorithm and cloud authentication certification. The detailed analysis shows it can resist the potential attacks. Besides, through performance evaluations, we further demonstrated the proposed scheme meets the efficiency need well.

REFERENCES

[1] A. Jain, L. Hong and S. Pankanti, “Biometric identification,” Communications of the ACM, vol. 43, no. 2, pp. 90-98, 2000.
[2] R. Allen, P. Sankar and S. Prabhakar, “Fingerprint identification technology,” Biometric Systems, pp. 22-61, 2005.
[3] J. de Mira, H. Neto, E. Neves, et al., “Biometric-oriented Iris Identification Based on Mathematical Morphology,” Journal of Signal Processing Systems, vol. 80, no. 2, pp. 181-195, 2015.
[4] S. Romdhani, V. Blanz and T. Vetter, “Face identification by fitting a 3d morphable model using linear shape and texture error functions,” in European Conference on Computer Vision, pp. 3-19, 2002.
[5] Y. Xiao, V. Rayi, B. Sun, X. Du, F. Hu, and M. Galloway, “A survey of key management schemes in wireless sensor networks,” Journal of Computer Communications, vol. 30, no. 11-12, pp. 2314-2341, 2007.
[6] 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, 2007.
[7] X. Du and H. H. Chen, “Security in wireless sensor networks,” IEEE Wireless Communications Magazine, vol. 15, no. 4, pp. 60-66, 2008.
[8] X. Hei, and X. Du, “Biometric-based two-level secure access control for implantable medical devices during emergency,” in Proc. of IEEE INFOCOM 2011, pp. 346-350, 2011.
[9] X. Hei, X. Du, J. Wu, and F. Hu, “Defending resource depletion attacks on implantable medical devices,” in Proc. of IEEE GLOBECOM 2010, pp. 1-5, 2010.
[10] M. Barni, T. Bianchi, D. Catalano, et al., “Privacy-preserving fingercode authentication,” in Proceedings of the 12th ACM workshop on Multimedia and security, pp. 231-240, 2010.
[11] M. Osadchy, B. Pinkas, A. Jarrous, et al., “SCiFI-a system for secure face identification,” in Security and Privacy (SP), 2010 IEEE Symposium on, pp. 239-254, 2010.
[12] D. Evans, Y. Huang, J. Katz, et al., “Efficient privacy-preserving biometric identification,” in Proceedings of the 17th conference Network and Distributed System Security Symposium, NDSS, 2011.
[13] J. Yuan and S. Yu, “Efficient privacy-preserving biometric identification in cloud computing,” in Proc. of IEEE INFOCOM 2013, pp. 2652-2660,2013.
[14] Q. Wang, S. Hu, K. Ren, et al., “CloudBI: Practical privacy-preserving outsourcing of biometric identification in the cloud,” in European Symposium on Research in Computer Security, pp. 186-205, 2015.
[15] Y. Zhu, Z. Wang and J. Wang, “Collusion-resisting secure nearest neighbor query over encrypted data in cloud,” In Quality of Service (IWQoS), 2016 IEEE/ACM 24th International Symposium on, pp. 1-6, 2016.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

L.Shruthi
M.Tech. SCHOLAR,
CSE Department
Malla Reddy College of Engineering

Abstract—As a side effect of increasingly popular social media, cyberbullying has emerged as a serious problem afflicting children, adolescents and young adults. Machine learning techniques make automatic detection of bullying messages in social media possi ble, and this could help to construct a healthy and safe social media environment. In this meaningful research area, one critical issue is robust and discriminative numerical representation learning of text messages. In this paper, we propose a new representation learning method to tackle this problem. Our method named Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) is developed via semantic extension of the popular deep learning model stacked denoising autoencoder. The semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique. Our proposed method is able to exploit the hidden feature structure of bullying information and learn a robust and discriminative representation of text. Comprehensive experiments on two public cyberbullying corpora (*Twitter* and *MySpace*) are conducted, and the results show that our proposed approaches outperform other baseline text representation learning methods.

Index Terms—Cyberbullying Detection, Text Mining, Representation Learning, Stacked Denoising Autoencoders, Word Embedding

1 INTRODUCTION

SOCIAL Media, as defined in [1], is ‘’a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.‘’ Via social media, people can enjoy enormous information, convenient communication experience and so on. However, social me- dia may have some side effects such as cyberbullying, which may have negative impacts on the life of people, especially
children and teenagers.

Cyberbullying can be defined as aggressive, intentional actions performed by an individual or a group of people via digital communication methods such as sending messages and posting comments against a victim. Different from tra- ditional bullying that usually occurs at school during face- to-face communication, cyberbullying on social media can take place anywhere at any time. For bullies, they are free to hurt their peers’ feelings because they do not need to face someone and can hide behind the Internet. For victims, they are easily exposed to harassment since all of us, especially youth, are constantly connected to Internet or social media. As reported in [2], cyberbullying victimization rate ranges from 10% to 40%. In the United States, approximately 43% of teenagers were ever bullied on social media [3]. The same as traditional bullying, cyberbullying has negative, insidious and sweeping impacts on children [4], [5], [6]. The outcomes for victims under cyberbullying may even be tragic such as the occurrence of self-injurious behaviour or suicides.

•

One way to address the cyberbullying problem is to automatically detect and promptlyreport bullying messages so that proper measures can be taken to prevent possi- ble tragedies. Previous works on computational studies of bullying have shown that natural language processing and machine learning are powerful tools to study bullying [7], [8]. Cyberbullying detection can be formulated as a supervised learning problem. A classifier is first trained on a cyberbullying corpus labeled by humans, and the learned classifier is then used to recognize a bullying message. Three kinds of information including text, user demography, and social network features are often used in cyberbullying detection [9]. Since the text content is the most reliable, our work here focuses on text-based cyberbullying detection. In the text-based cyberbullying detection, the first and also critical step is the numerical representation learning for text messages. In fact, representation learning of text is extensively studied in text mining, information retrieval and natural language processing (NLP). Bag-of-words (BoW) model is one commonly used model that each dimension corresponds to a term. Latent Semantic Analysis (LSA) and topic models are another popular text representation models, which are both based on BoW models. By mapping text units into fixed-length vectors, the learned represen- tation can be further processed for numerous language processing tasks. Therefore, the useful representation should discover the meaning behind text units. In cyberbullying detection, the numerical representation for Internet mes- sages should be robust and discriminative. Since messages on social media are often very short and contain a lot of informal language and misspellings, robust representations for these messages are required to reduce their ambigui-
ty. Even worse, the lack of sufficient high-quality training

ISBN 978-93-88808-61-3

data, i.e., data sparsity make the issue more challenging. Firstly, labeling data is labor intensive and time consuming. Secondly, cyberbullying is hard to describe and judge from a third view due to its intrinsic ambiguities. Thirdly, due to protection of Internet users and privacy issues, only a small portion of messages are left on the Internet, and most bullying posts are deleted. As a result, the trained classifier may not generalize well on testing messages that contain nonactivated but discriminative features. The goal of this present study is to develop methods that can learn robust and discriminative representations to tackle the above prob- lems in cyberbullying detection.

Some approaches have been proposed to tackle these problems by incorporating expert knowledge into feature learning. Yin et.al proposed to combine BoW features, senti- ment features and contextual features to train a support vec- tor machine for online harassment detection [10]. Dinakar et.al utilized label specific features to extend the general features, where the label specific features are learned by Linear Discriminative Analysis [11]. In addition, common sense knowledge was also applied. Nahar et.al presented a weighted TF-IDF scheme via scaling bullying-like features by a factor of two [12]. Besides content-based information, Maral et.al proposed to apply users' information, such as gender and history messages, and context information as extra features [13], [14]. But a major limitation of these approaches is that the learned feature space still relies on the BoW assumption and may not be robust. In addition, the performance of these approaches rely on the quality of hand-crafted features, which require extensive domain knowledge.

In this paper, we investigate one deep learning method named stacked denoising autoencoder (SDA) [15]. SDA stacks several denoising autoencoders and concatenates the output of each layer as the learned representation. Each denoising autoencoder in SDA is trained to recover the input data from a corrupted version of it. The input is corrupted by randomly setting some of the input to zero, which is called dropout noise. This denoising process helps the autoencoders to learn robust representation. In addition, each autoencoder layer is intended to learn an increasingly abstract representation of the input [16]. In this paper, we develop a new text representation model based on a variant of SDA: marginalized stacked denoising autoencoders (mS- DA) [17], which adopts linear instead of nonlinear projection to accelerate training and marginalizes infinite noise distri- bution in order to learn more robust representations. We utilize semantic information to expand mSDA and develop Semantic-enhanced Marginalized Stacked Denoising Au- toencoders (smSDA). The semantic information consists of bullying words. An automatic extraction of bullying words based on word embeddings is proposed so that the involved human labor can be reduced. During training of smSDA, we attempt to reconstruct bullying features from other normal words by discovering the latent structure, i.e. correlation, between bullying and normal words. The intuition behind this idea is that some bullying messages do not contain bullying words. The correlation information discovered by smSDA helps to reconstruct bullying features from normal words, and this in turn facilitates detection of bullying messages without containing bullying words. For example, there is a strong correlation between bullying word *fuck* and normal word *off* since they often occur together. If bullying messages do not contain such obvious bullying features, such as *fuck* is often misspelled as *fck*, the correlation may help to reconstruct the bullying features from normal ones so that the bullying message can be detected. It should be noted that introducing dropout noise has the effects of enlarging the size of the dataset, including training data size, which helps alleviate the data sparsity problem. In addition, L1 regularization of the projection matrix is added to the objective function of each autoencoder layer in our model to enforce the sparstiy of projection matrix, and this in turn facilitates the discovery of the most relevant terms for reconstructing bullying terms. The main contributions of our work can be summarized as follows:

* Our proposed Semantic-enhanced Marginalized S- tacked Denoising Autoencoder is able to learn ro- bust features from BoW representation in an effi- cient and effective way. These robust features are learned by reconstructing original input from cor- rupted (i.e., missing) ones. The new feature space can improve the performance of cyberbullying de- tection even with a small labeled training corpus.
* Semantic information is incorporated into the re- construction process via the designing of semantic dropout noises and imposing sparsity constraints on mapping matrix. In our framework, high-quality semantic information, i.e., bullying words, can be extracted automatically through word embeddings. Finally, these specialized modifications make the new feature space more discriminative and this in turn facilitates bullying detection.
* Comprehensive experiments on real-data sets have verified the performance of our proposed model.

This paper is organized as follows. In Section 2, some re- lated work is introduced. The proposed Semantic-enhanced Marginalized Stacked Denoising Auto-encoder for cyber- bullying detection is presented in Section 3. In Section 4, experimental results on several collections of cyberbullying data are illustrated. Finally, concluding remarks are provid- ed in Section 5.

2 Related Work

This work aims to learn a robust and discriminative text rep- resentation for cyberbullying detection. Text representation and automatic cyberbullying detection are both related to our work. In the following, we briefly review the previous work in these two areas.

Text Representation Learning

In text mining, information retrieval and natural language processing, effective numerical representation of linguistic units is a key issue. The Bag-of-words (BoW) model is the most classical text representation and the cornerstone of some states- of-arts models including Latent Semantic Analysis (LSA) [18] and topic models [19], [20]. BoW model represents a document in a textual corpus using a vector of real numbers indicating the occurrence of words in the document. Although BoW model has proven to be efficient

and effective, the representation is often very sparse. To address this problem, LSA applies Singular ValueDecompo- sition (SVD) on the word-document matrix for BoW model to derive a low-rank approximation. Each new feature is a linear combination of all original features to alleviate the sparsity problem. Topic models, including Probabilistic La- tent Semantic Analysis [21] and Latent Dirichlet Allocation [20], are also proposed. The basic idea behind topic models is that word choice in a document will be influenced by the topic of the document probabilistically. Topic models try to define the generation process of each word occurred in a document.

Similar to the approaches aforementioned, our proposed approach takes the BoW representation as the input. How- ever, our approach has some distinct merits. Firstly, the mul- ti-layers and non-linearity of our model can ensure a deep learning architecture for text representation, which has been proven to be effective for learning high-level features [22]. Second, the applied dropout noise can make the learned representation more robust. Third, specific to cyberbullying detection, our method employs the semantic information, including bullying words and sparsity constraint imposed on mapping matrix in each layer and this will in turn pro- duce more discriminative representation.

Cyberbullying Detection

With the increasing popularity of social media in recent years, cyberbullying has emerged as a serious problem afflicting children and young adults. Previous studies of cyberbullying focused on extensive surveys and its psy- chological effects on victims, and were mainly conducted by social scientists and psychologists [6], [23], [24], [25]. Although these efforts facilitate our understanding for cyberbullying, the psychological science approach based on personal surveys is very time-consuming and may not be suitable for automatic detection of cyberbullying. Since machine learning is gaining increased popularity in recent years, the computational study of cyberbullying has at- tracted the interest of researchers. Several research areas including topic detection and affective analysis are closely related to cyberbullying detection. Owing to their efforts, automatic cyberbullying detection is becoming possible. In machine learning-based cyberbullying detection, there are two issues: 1) text representation learning to transform each post/message into a numerical vector and 2) classifier train- ing. Xu et.al presented several off-the-shelf NLP solutions including BoW models, LSA and LDA for representation learning to capture bullying signals in social media [8]. As an introductory work, they did not develop specialized models for cyberbullying detection. Yin et.al proposed to combine BoW features, sentiment feature and contextual features to train a classifier for detecting possible harassing posts [10]. The introduction of the sentiment and contextual features has been proven to be effective. Dinakar et.al used Linear Discriminative Analysis to learn label specific features and combine them with BoW features to train a classifier [11]. The performance of label-specific features largely depends on the size of training corpus. In addition, they need to construct a bullyspace knowledge base to boost the performance of natural language processing methods. Although the incorporation of knowledge base can achieve a performance improvement, the construction of a complete and general one is labor-consuming. Nahar et.al proposed to scale bullying words by a factor of two in the original BoW features [12]. The motivation behind this work is quit similar to that of our model to enhance bullying features. However, the scaling operation in [12] is quite arbitrary.Ptaszynski et.al searched sophisticated patterns in a brute-force way [26]. The weights for each extracted pattern need to be calcu- lated based on annotated training corpus, and thus the per- formance may not be guaranteed if the training corpus has a limited size. Besides content-based information, Maral et.al also employ users' information, such as gender and history messages, and context information as extra features [13], [14]. Huang et.al also considered social network features to learn the features for cyberbullying detection [9]. The shared deficiency among these forementioned approaches is constructed text features are still from BoW representation, which has been criticized for its inherent over-sparsity and failure to capture semantic structure [18], [19], [20]. Differ- ent from these approaches, our proposed model can learn robust features by reconstructing the original data from corrupted data and introduce semantic corruption noise and sparsity mapping matrix to explore the feature structure which are predictive of the existence of bullying so that the learned representation can be discriminative.

3 Semantic-Enhanced Marginalized Stacked Denoising Auto-encoder

We first introduce notations used in our paper. Let $D = \{w_1, \ldots, w_d\}$ be the dictionary covering all the words exist- ing in the text corpus. We represent each message using a BoW vector $\mathbf{x} \in \mathrm{R}^d$. Then, the whole corpus can be denoted as a matrix: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathrm{R}^{d \times n}$, where n is the number of available posts.

We next briefly review the marginalized stacked de- noising auto-encoder and present our proposed Semantic- enhanced Marginalized Stacked Denoising Auto-Encoder.

Marginalized Stacked Denoising Auto-encoder

Chen et.al proposed a modified version of Stacked Denois- ing Auto-encoder that employs a linear instead of a non- linear projection so as to obtain a closed-form solution [17]. The basic idea behind denoising auto-encoder is to recon- struct the original input from a corrupted one $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ with the goal of obtaining robust representation.

Marginalized Denoising Auto-encoder: In this mod- el, denoising auto-encoder attempts to reconstruct original data using the corrupted data via a linear projection. The projection matrix can be learned as:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{W}\tilde{\mathbf{x}}_i\|_2 \quad (1)$$

where $\mathbf{W} \in \mathrm{R}^{d \times d}$. For simplicity, we can write Eq. (1) in matrix form as:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}} \frac{1}{2n} tr\left((\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}})^{\mathrm{T}}(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}})\right) \quad (2)$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]$ is the corrupted version of $\mathbf{X}$. It is easily shown that Eq. (2) is an ordinary least square problem having a closed-form solution:

$$\mathbf{W} = \mathbf{P}\mathbf{Q}^{-1} \tag{3}$$

where $\mathbf{P} = \mathbf{X}\tilde{\mathbf{X}}^{\mathrm{T}}$ and $\mathbf{Q} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}$. In fact, this corruption can be marginalized over the noise distribution [17]. The more corruptions we take in the denoising auto-encoder, the more robust transformation can be learned. Therefore, the best choice is using infinite versions of corrupted data. If the data corpus is corrupted infinite times, the matrix P and Q are converged to their corresponding expectation, andEq.
(3) can be formulated as:

$$\mathbf{W} = E[\mathbf{P}]\,E[\mathbf{Q}]^{-1} \tag{4}$$

where $E[\mathbf{P}] = \sum_{n} E\,\mathbf{x}_i\tilde{\mathbf{x}}^{\mathrm{T}}$ and $E[\mathbf{Q}] = \sum_{n=1} E\,\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}^{\mathrm{T}}$. These expected matrices can be computed based on noise distribution. In [17], dropout noise is adopted to corrupt data samples by setting a feature to zero with a probability p. Assuming the scatter matrix of the original data samples is denoted as $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$, the expected matrices can be computed as:

$$E[\mathbf{Q}]_{i,j} = \begin{cases} (1-p)^2\mathbf{S}_{i,j} & \text{if } i \neq j, \\ (1-p)\mathbf{S}_{i,j} & \text{if } i = j. \end{cases} \tag{5}$$

and

$$E[\mathbf{P}]_{i,j} = (1-p)\mathbf{S}_{i,j} \tag{6}$$

where i and j denotes the indices of features. It can be seen that it is very efficient to compute $\mathbf{W}$ by marginalizing dropout noise in denoising auto-encoder. After the mapping weights $\mathbf{W}$ are computed, a nonlinear squashing function, such as a hyperbolic tangent function, can be applied to de- rive the output of the marginalized denoisingauto-encoder:

$$\mathbf{H} = \tanh(\mathbf{W}\mathbf{X}) \tag{7}$$

Stacking Structure: Chen et.al [17] also proposed to apply stacking structures on marginalized denoising au-
toencoder, in which the output of the $(k-1)^{th}$ layer is fed as the input into the k^{th} layer. If we define the output of the k^{th} mDA as $\mathbf{H}_k$ and the original input as $\mathbf{H}_0$ respectively, the mapping between two consecutive layers is given as:

$$\mathbf{H}_k = \tanh(\mathbf{W}_k\mathbf{H}_{k-1}) \tag{8}$$

where $\mathbf{W}_k$ denotes the mapping in k^{th} layer. The model training can be done greedily layer by layer. This means that the mapping weights $\mathbf{W}_k$ is learned in a closed-form to
reconstruct the output of $(k-1)^{th}$mDA layer from its marginalized corruptions, as shown in Eq. (4). If the number of layers is set to L, the final representation for input data $\mathbf{X}$ is the concatenation of the uncorrupted original input and outputs of all layers as follows:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_L \end{bmatrix} \tag{9}$$

where $\mathbf{Z} \in \mathrm{R}^{d(L+1)\times n}$. Each column of $\mathbf{Z}$ represents the final representation of each individual data sample.

Semantic Enhancement for mSDA

The advantage of corrupting the original input in mSDA can be explained by feature co-occurrence statistics. The co- occurrence information is able to derive a robust feature rep- resentation under an unsupervised learning framework, and this also motivates other state-of-the-art text feature learning methods such as Latent Semantic Analysis and topicmodels [18], [20]. As shown in Figure 1. (a), a denoising auto- encoder is trained to reconstruct these removed features values from the rest uncorrupted ones. Thus, the learned mapping matrix W is able to capture correlation between these removed features and other features. It is shown that
the learned representation is robust and can be regarded as a high level concept feature since the correlation informa-
tion is invariant to domain-specific vocabularies. We next describe how to extend mSDA for cyberbullying detection. The major modifications include semantic droupout noise and sparse mapping constraints.

Semantic Dropout Noise

The dropout noise adopted in mSDA is an uniform distribution, where each feature has the same probability to be removed. In cyberbullying detection, most bullying posts contain bullying words such as profanity words and foul languages. These bullying words are very predictive of the existence of cyberbullying. However, a direct use of these bullying features may not achieve good performance because these words only account for a small portion of the whole vocabulary and these vulgar words are only onekind of discriminative features for bullying [10], [26]. In other way, we can explore these cyberbullying words by using a different dropout noise that features corresponding to bullying words have a larger probability of corruption than other features. The imposed large probability on bullying words emphasizes the correlation between bullyingfeatures and normal ones. This kind of dropout noise can be denoted as semantic dropout noise, because semantic information is used to design dropout structure.

As shown in Figure 1. (b), the correlation between fea- tures can enable other normal words to predict bullying labels. Considering a simple but intuitive example, "Leave him alone, he is just a chink"[1], which is obviously a bullying message. However, the classifier will set the weight of the discriminative word "chink" to zero, if the small sized training corpus does not cover it. Our proposed smSDA can deal with the problem by learning a robust feature represen- tation, which is a high level concept representation. In the learned representation, the word "chink" are reconstruct- ed by context words co-occurring with the specific word ("chink") and the context words may be shared by other bullying words contained in training corpus. Therefore, the correlation explored by this auto-encoder structure enables the subsequent classifier to learn the discriminative word and improve the classification performance. In addition, the semantic dropout noise exploits the correlation between

1. *"Chink (also chinki, chinky, chinkie) is an English ethnic slur usually referring to a person of Chinese or East Asian ethnicity"* from Wikipedia

bullying features and normal features better and hence, facilitates cyberbullying detection.

Due to the introduced semantic dropout noise, the ex- pected matrices: $E[\mathbf{P}]$ and $E[\mathbf{Q}]$ will be computed slightly different from Eqs. (5) and (6). Assuming we have an available bullying words list and the correspondingfeatures set Z_b, the semantic dropout noise can be described as the following probability density function (PDF):

$$PDF = \begin{cases} p(\tilde{x}_d = 0) = p_n & \text{if } d \notin Z_b, \\ p(\tilde{x}_d = x_d) = 1 - p_n & \text{if } d \notin Z_b, \\ p(\tilde{x}_d = 0) = p_b & \text{if } d \in Z, \\ p(\tilde{x}_d = x_d) = 1 - p_b & \text{if } d \in Z_b, \end{cases} \quad (10)$$

where d denotes the feature set. Then these two marginal- ized matrices can be computed as:

(16)

where λ is a regularization parameter that controls the

$$E[\mathbf{Q}]_{i,j} = \begin{cases} (1-p_n)\mathbf{S}_{i,j} & \text{if } i = j \;\&\; i \notin Z_b, \\ (1-p_b)^2\mathbf{S}_{i,j} & \text{if } i \neq j \;\&\; \{i,j\} \cap Z_b = \emptyset, \\ (1-p_b)(1-p_n)\mathbf{S}_{i,j} & \text{if } \{i,j\} \notin Z_b \;\&\; \{i,j\} \cap Z_b \neq \emptyset, \\ (1-p_b)^2\mathbf{S}_{i,j} & \text{if } i \neq j \;\&\; \{i,j\} \in Z_b, \\ (1-p_b)\mathbf{S}_{i,j} & \text{if } i = j \;\&\; i \in Z_b \end{cases} \quad (11)$$

and

$$E[\mathbf{P}]_{i,j} = \begin{cases} (1-p_n)\mathbf{S}_{i,j} & \text{if } j \cap Z_b = \emptyset, \\ (1-p_b)\mathbf{S}_{i,j} & \text{if } j \cap Z_b \neq \emptyset. \end{cases} \quad (12)$$

where p_b and p_n are the probabilities of bullying features and normal features to be set to zero respectively, and $p_b > p_n$. Here, p_b and p_n are both tunable hyperparameters for our proposed smSDA.

Unbiased Semantic Dropout Noise As shown in Eq . (10), the corrupted data is biased, i.e., $E[\mathbf{X}] \neq E[\tilde{\mathbf{X}}]$. Here, we modified Eq. (10) to achieve an unbiased noise as follows:

$$PDF^{unbiased} = \begin{cases} p(\tilde{x}_d = 0) = p_n & \text{if } d \notin Z_b, \\ p(\tilde{x}_d = \frac{x_d}{1-p_n}) = 1 - p_n & \text{if } d \notin Z_b, \\ p(\tilde{x}_d = 0) = p_b & \text{if } d \in Z_b, \\ p(\tilde{x}_d = \frac{x_d}{1-p_b}) = 1 - p_b & \text{if } d \in Z_b, \end{cases} \quad (13)$$

It can be easily shown that under such a noise distribution, the corrupted data is unbiased now. These two marginalized matrices are re-formulated as:

$$E[\mathbf{Q}]^{unbiased}_{i,j} = \begin{cases} \frac{1}{1-p_n}\mathbf{S}_{i,j} & \text{if } i = j \;\&\; i \notin Z, \\ \frac{1}{(1-p_n)(1-p_b)}\mathbf{S}_{i,j} & \text{if } i = j \;\&\; i \in Z_b, \\ \mathbf{S}_{i,j} & \text{if } i \neq j. \end{cases} \quad (14)$$

and

$$E[\mathbf{P}]^{unbiased}_{i,j} = \mathbf{S}_{i,j} \quad (15)$$

These two computed matrices will then be used to learn the mapping in each layer in our proposed smSDA.

Sparsity Constraints

In mSDA, the mapping matrix $\mathbf{W}$ is learned to reconstruc- t removed features from other uncorrupted features and hence is able to capture the feature correlation information. Here, we inject the sparsity constraints on the mapping weights $\mathbf{W}$ so that each row has a small number of nonzero elements. This sparsity constraint is quite intuitive because one word is only related to a small portion of vocabulary instead of the whole vocabulary. In our proposed smSDA, the sparsity constraint is realized by the incorporation of L1 regularization term into the objective function as in the lasso problem [27]. The optimization function for each layer in smSDA is given as follows:

$$\mathbf{W} = \arg\min_{w} \frac{1}{2n} tr\,(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}})^T(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}}) + \lambda\|\mathbf{W}\|_1$$

sparsity of $\mathbf{W}$. The larger the λ is, the sparser the mapping matrix $\mathbf{W}$ is. The solution to Eq. (16) is a very mature math problem: sparse least squares optimization, which has several effective and efficient computation methods [28], [29], [30]. Here, we adopt a method called Iterated Ridge Regression, which has been proven to be very efficient [30]. The method firstly introduces an approximation:

$$\| \approx \frac{\mathbf{w}_i^T\mathbf{w}_i}{\|\mathbf{w}_i\|_1} \quad (17)$$

where $\mathbf{w}_i$ denotes the i-th row in the whole matrix $\mathbf{W}$. By substituting this approximation Eq. (17) into the objective function Eq. (16), we yield an formulation similar to a Ridge Regression Problem [31], and the iteration steps to solve $\mathbf{W}$ is given as:

$$\mathbf{W}_k = \left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \lambda\,\mathrm{diag}(|\mathbf{W}_{k-1}|)\right)^{-1}\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\right) \quad (18)$$

where diag denotes the diagonal elements of a matrix, $\mathbf{W}_k$ and $\mathbf{W}_{k-1}$ denote the current step and the previous step estimations for mapping matrix $\mathbf{W}$ respectively. It is clear that the Eq. (18) can be easily formulated when the noise distribution is marginalized. Similar to Eq. (4), Eq. (18) can be written as:

$$\mathbf{W}_k = E[\mathbf{P}]\left(E[\mathbf{Q}] + \lambda\,\mathrm{diag}(|\mathbf{W}_{k-1}|)^{-1}\right) \quad (19)$$

To speed up the convergence process, the initialization for $\mathbf{W}$ can be set to the L2 penalized solution for Eq. (2) as follows:

$$\mathbf{W}_0 = E[\mathbf{P}]\left(E[\mathbf{Q}] + \lambda\mathbf{I}\right) \quad (20)$$

where $\mathbf{I}$ is an identify matrix. It can be shown that this itera- tion procedure can also marginalize the noise distribution easily, which can ensure an efficient and stable mapping learning.

ISBN 978-93-88808-61-3

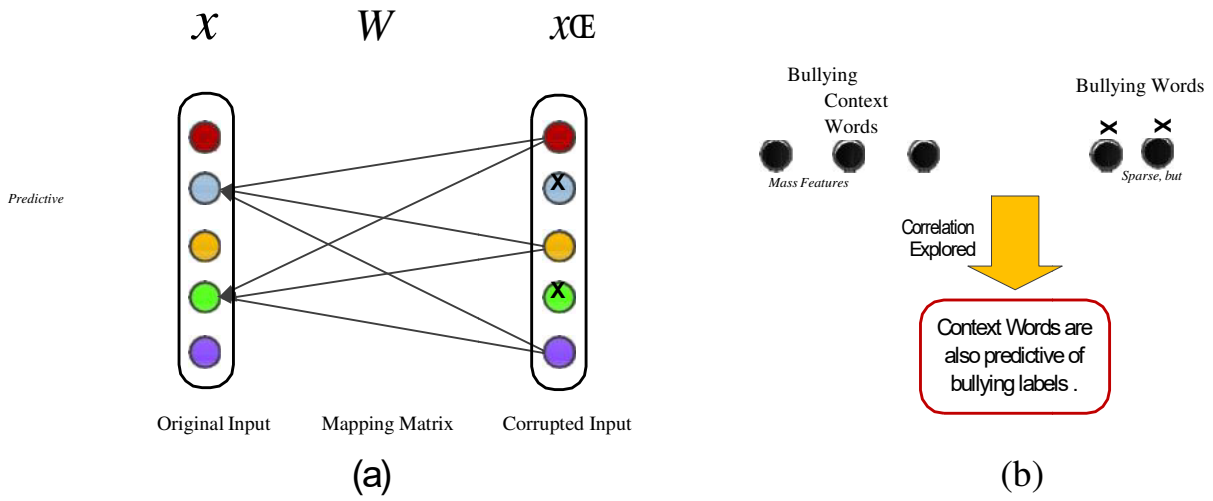


Fig. 1. Illustration of Motivations behind smSDA. In Figure 1(a), the cross symbol denotes that its corresponding feature is corrupted, i.e., turned off.

Construction of Bullying Feature Set

As analyzed above, the bullying features play an important role and should be chosen properly. In the following, the steps for constructing bullying feature set Z_b are given, in which the first layer and the other layers are addressed separately. For the first layer, expert knowledge and word embeddings are used. For the other layers, discriminative feature selection is conducted.

Layer One: firstly, we build a list of words with negative affective, including swear words and dirty words. Then, we compare the word list with the BoW features of our own corpus, and regard the intersections as bullying features. However, it is possible that expert knowledge is limited and does not reflect the current usage and style of cyberlanguage. Therefore, we expand the list of pre-defined insulting words, i.e. *insulting seeds*, based on word embeddingsas follows:

Word embeddings use real-valued and low-dimensional vectors to represent semantics of words [32], [33] . The well-trained word embeddings lie in a vector space where similar words are placed close to each other. In addition, the cosine similarity between word embeddings is able to quantify the semantic similarity between words. Considering the Interent messages are our interested corpus, we utilize a well-trained word2vec model on a large-scale twittercorpus containing 400 million tweets [34]. A visualization of some word embeddings after dimensionality reduction (PCA) is shown in Figure 2. It is observed that curse words form distinct clusters, which are also far away from normal words. Even insulting words are located at different regions due to different word usages and insulting expressions. In addition, since the word embeddings adopted here are trained in a large scale corpus from Twitter, the similarity captured by word embeddings can represent the specific language pattern. For example, the embedding of the misspelled word *fck* is close to the embedding of *fuck* so that the word *fck* can be automatically extracted based on word embeddings.

We extend the pre-defined *insulting seeds* based onword embeddings. For each insulting seed, similar words are extracted if their cosine similarities with insult seed exceed a predefined threshold. For bigram $w_l w_r$, we simply use an additive model to derive the corresponding embedding as follows:

$$\mathbf{v}(w_l w_r) = \mathbf{v}(w_l) + \mathbf{v}(w_r) \quad (21)$$

Finally, the constructed bullying features are used to train the first layer in our proposed smSDA. It includes two parts: one is the original *insulting seeds* based on domain knowledge and the other is the extended bullying words via word embeddings. The length of Z_b is k.

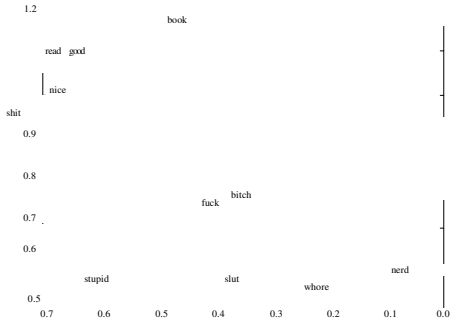


Fig. 2. Two dimensional visualization of our used word embeddings via PCA. Displayed terms include both bullying ones and normal ones. It shows that similar words are nearby vectors.

Subsequent Layers: we perform feature selection using Fisher score to select ''bullying'' features. Fisher score is an univariate metric reflecting the discriminative power of a feature [35], [36]. For the r^{th} feature, the corresponding Fisher score can be computed based on training data with

$$\frac{\sum_{c} n_i (\mu_i - \mu)^2}{\sum_{i=1}^{c} n_i \sigma_i^2} \quad (22)$$

where c denotes the number of classes and n_i represent the number of data in class i. μ and μ_i denote the mean of entire data and class i for the r^{th} feature, and σ_i is the variance of class i on r^{th} feature. After Fisher scores are estimated, fea- tures with top k scores are selected as ''bullying'' features, where ''bullying'' is generalized as discriminative.

smSDA for Cyberbullying Detection

In section 3.3, we propose the Semantic-enhanced Marginal- ized Stacked Denoising Auto-encoder (smSDA). In this subsection, we describe how to leverage it for cyberbully- ing detection. smSDA provides robust and discriminative representations The learned numerical representations can then be fed into Support Vector Machine (SVM). In the new space, due to the captured feature correlation and semantic information, the SVM, even trained in a small size of training corpus, is able to achieve a good performance on testing documents(this will be verified in the following experiments). The detailed steps of our model are provided below:

Assuming the first n_l posts are labeled and the corre- sponding vector of binary labels is $\mathbf{y} = y_1, \ldots, y_{n_l}$. The binarylabel 1 or 0 indicates the postis oris not acyberbully- ing one. Here, n_l n, which means the labeled posts have a small size. The bullying feature set Z_b is constructed in a layer-wise way. Based on prior knowledge, we construct are selected as *insulting seeds*. The insulting seeds are then expanded and refined automatically via word embeddings, which defines the bullying features Z_b for layer one. The experiments in Section 4 will show that the construction of the set Z_b is very simple and efficient with litter human labor. For the subsequent layers, after obtaining the output of each layer, the set Z_b is updated using feature ranking with Fish score according to Eq. (22).

Based on predefined dropout probabilities for bullying features and other normal features p_b and p_n and the bullying feature set Z_b, we compute these two expected matrices $E[\mathbf{P}]$ and $E[\mathbf{Q}]$ according to Eqs. (12) and (11), if the semantic dropout noise is adopted. When it comes to the unbiased semantic dropout noise, Eqs. (14) and (15) instead of Eqs. (12) and (11) are used to compute these two expected matrices. Then, we iteratively perform Eq. (21) for T_{max} times, where the initial value for $\mathbf{W}$ is calculated based on Eq. (20). When the mapping matrix is learned, the output of each layer is given according to Eq. (8). Due to the stacking structure, the output of L layers and the initial input are concatenated together to form the final representation $\mathbf{Z} \in R^{d(L+1)\times n}$ following Eq. (9). It is clear that the new space has a dimension of $(L + 1)d$. A linear SVM [37] is trained on the training corpus, i.e. the first n_l columns in $\mathbf{Z}$ and tested on the rest data samples.

Merits of smSDA

Some important merits of our proposed approach are sum- marized as follows:

1) Most cyberbullying detection methods rely on the BoW model. Due to the sparsity problems of both data and features, the classifier may not be trained very well. Stacked densoing autoencoder (SDA), as an unsupervised representation learning method, is able to learn a robust feature space. In SDA, the feature correlation is explored by the reconstruction of corrupted data. The learned robust feature rep- resentation can then boost the training of classifier and finally improve the classification accuracy. In addition, the corruption of data in SDA actually generates artificial data to expand data size, which alleviate the small size problem of training data.
2) For cyberbullying problem, we design semantic dropout noise to emphasize bullying features in the new feature space, and the yielded new represen- tation is thus more discriminative for cyberbullying detection.
3) The sparsity constraint is injected into the solution of mapping matrix $\mathbf{W}$ for each layer, considering each word is only correlated to a small portion of the whole vocabulary. We formulate the solution for the mapping weights $\mathbf{W}$ as an Iterated Ridge Regres- sion problem, in which the semantic dropout noise distribution can be easily marginalized to ensurethe efficient training of our proposed smSDA.
4) Based on word embeddings, bullying features can be extracted automatically. In addition, the possible limitation of expert knowledge can be alleviated by the use of word embedding.

4 EXPERIMENTS

In this section, we evaluate our proposed semantic- enhanced marginalized stacked denoising auto-encoder (smSDA) with two public real-world cyberbullying corpora. We start by describing the adopted corpora and experimen- tal setup. Experimental results are then compared with other baseline methods to test the performance of our approach. At last, we provide a detailed analysis to explain the good performance of our method.

Descriptions of Datasets

Two datasets are used here. One is from *Twitter* and another is from *MySpace* groups. The details of these two datasets are described below:

***Twitter* Dataset**: *Twitter* is ''a real-time information net- work that connects you to the latest stories, ideas, opin- ions and news about what you find interesting'' (https: //about.twitter.com/). Registered users can read and post tweets, which are defined as the messages posted on *Twitter* with a maximum length of 140 characters.

The *Twitter* dataset is composed of tweets crawled by the public *Twitter* stream API through two steps. In Step 1, keywords starting with ″bull″ including ″bully″, ″bullied″ and ″bullying″ are used as queries in Twitter to preselect some tweets that potentially contain bullying contents. Re- tweets are removed by excluding tweets containing the acronym ''RT''. In Step 2, the selected tweets are manually labeled as bullying trace or non-bullying trace based on the contents of the tweets. 7321 tweets are randomly sampled from the whole tweets collections from August 6, 2011 to

ISBN 978-93-88808-61-3

August 31, 2011 and manually labeled[2]. It should be pointed out here that labeling is based on bullying traces. A bullying trace is defined as the response of participants to their bullying experience. Bullying traces include not only mes- sages about direct bullying attack, but also messages about reporting a bullying experience, revealing self as a victim et. al. Therefore, bullying traces far exceed the incidents of cyberbullying. Automatic detection of bullying traces are valuable for cyberbullying research [38]. Some examples of bullying traces are shown in Figure 3. To preprocess these tweets, a tokenizer is applied without any stemming or stopword removal operations. In addition, some special characters including user mentions, URLS and so on are replaced by predefined characters, respectively. The features are composed of unigrams and bigrams that should appear at least twice and the details of preprocessing can be found in [8]. The statistics of this dataset can be found in Table 1. ***MySpace* Dataset**: *MySpace* is another web2.0 social net- working website. The registered accounts are allowed to view pictures, read chat and check other peoples' profile information.

The *MySpace* dataset is crawled from *MySpace* groups. Each group consists of several posts by different users, which can be regarded as a conversation about one topic. Due to the interactive nature behind cyberbullying, each data sample is defined as a window of 10 consecutive posts and the windows are moved one post by one post so that we got multiple windows [39]. Then, three people labeled the data for the existence of bullying content independently. To be objective, an instance is labeled as cyberbullying only if at least 2 out of 3 coders identify bullying content in the windows of posts. The raw text for these data, as XML files, have been kindly provided by Kontostathis et.al[3]. The XML files contain information about the posts, such as post text, post data, and users' information, which are put into 11 packets. Some posts in *MySpace* are shown in Figure 4.

Here, we focus on content-based mining, and hence, we only extract and preprocess the posts' text. The preprocessing steps of the *MySpace* raw text includetokenization, dele- tion of punctuation and special characters. The unigrams and bigrams features are adopted here. The threshold for negligible low-frequency terms is set to 20, considering one post occurred in a long conversation will occur in at least ten windows. The details of this dataset is shown in Table 1. Since there were no standard splits of training vs. test datasets in our adopted *Twitter* and *MySpace* corpora, we need to define the training and testing datasets. As analyzed above that the lack of labeled training corpus hinders the development of automatic cyberbullying detection, the sizes of training corpus are all controlled to be very small in our experiments. For *Twitter* dataset, we randomly select 800 instances, which accounts for 12% of the whole corpus, as the training data and the rest data samples are used as testing data. To reduce variance, the process is repeated ten times so that we can have ten *sub*-datasets from *Twitter* data. For *MySpace* dataset, we also randomly pick 400 data samples as the training corpus and use the rest data for testing. The process is repeated ten times to generate ten *sub*-datasets constructed from *MySpace* data. Finally, we have twenty *sub*-datasets, in which ten datasets are from *Twitter* corpus and another ten datasets are from *MySpace* corpus.

2. The dataset: **bullyingV3.0**, has been kindly provided at http:// research.cs.wisc.edu/bullying/data.html

3. The dataset: **MySpace Group**, has been kindly provided at http: //www.chatcoder.com/DataDownload

TABLE 1
Statistical Properties of the two datasets.

Statistics	*Twitter*	*MySpace*
Feature No.	4413	4240
Sample No.	7321	1539
Bullying Instances	2102	398

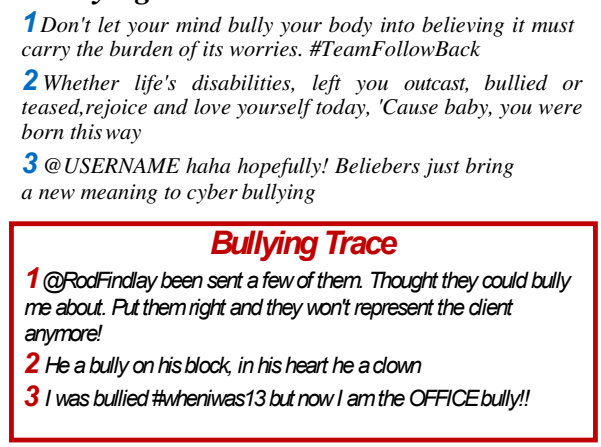


Fig. 3. Some Examples from *Twitter* Datasets. Three of them are non-bullying traces. And the other three are bullying traces.

Experimental Setup

Here, we experimentally evaluate our smSDA on two cyber-bullying detection corpora. The following methods will be compared.

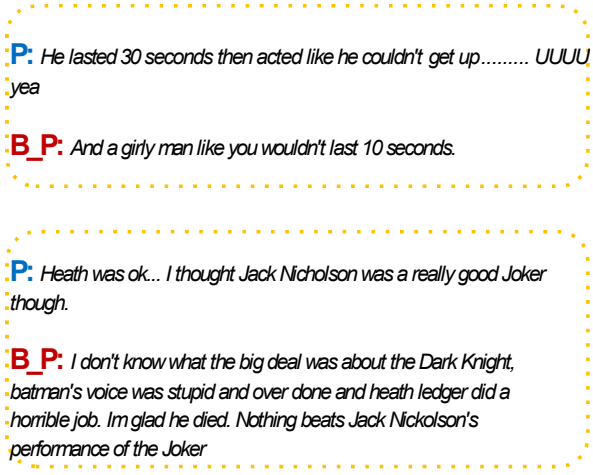


Fig. 4. Some Examples from *MySpace* Datasets. Two Conversions are Displayed and each one includes a normal post (P) and a bullyingpost ($B P$) .

ISBN 978-93-88808-61-3

* BWM: Bullying word matching. If the message con- tains at least one of our defined bullying words, it will be classified as bullying.
* BoW Model: the raw BoW features are directly fed into the classifier.
* Semantic-enhanced BoW Model: This approach is referred in [12]. Following the original setting, we scale the bullying features by a factor of 2.
* LSA: Latent Semantic Analysis [18].
* LDA: Latent Dirchilet Allocation [20]. Our implementation of LDA is based on *Gensim*[4].
* mSDA: marginalized stacked denoising autoen- coder [17].
* smSDA and smSDA$_u$: semantic-enhanced marginal- ized denoising autoencoder that utilizes semantic dropout noise and unbiased one, respectively.

For LSA and LDA, the number of latent topics are both set to 100. In LDA, we set hyperparameter α for document topic multinomial and hyperparameter η for word topic multinomial to 1 and 0.01, respectively. For mSDA[5], the noise intensity is set to 0.5 and the number of layers for *Tweets* and *MySpace* datasets are both set to 2. Here, the number of layers is only set to be a moderate number instead of a large one, considering a large final dimension will impose a computational burden on the subsequent classifier training.

For our proposed methods including smSDA and smSDA$_u$: the noise intensity and the number of layers are set to the same values as in mSDA to give a fair comparison. The bullying noise intensity is set to 0.8, which is larger than 0.5. The hyperparameters λ that controls the sparsity of the transformation matrix are set to 1 for all layers. The number of iteration step for solving lasso problems is set to
20. To construct the bullying features Z_b for the first layer,

whose word cloud visualization is shown in Figure 5. The intersections between BoW features of our own corpus and the predefined bullying word list are firstly obtained. Then, as described in 3.3, they are extended and refined based on word embeddings to form the final bullying features. The threshold for cosine similarity is set to 0.8. The word cloud

visualizations for the final bullying features in *Twitter* and *MySpace* datasets are shown in Figures 6 and 7, respective- ly. The bullying features used in Semantic-enhanced BoW Model are the same as those in smSDA.

Linear SVM [37] is then applied to the new feature space generated by the above mentioned approaches. In linear SVM, we search the best regularization parameter C from
$\{0.0001, 0.001, 0.01, 0.1, 1\}$. To evaluate the performance of these methods on binary classification, classification accuracy is employed. Considering both datasets have the class imbalance problem, we also introduce F1-Score, which is a balance between precision and recall, to evaluate the performance of all compared approaches.

4. https://radimrehurek.com/gensim/index.html
5. The code has been kindly provided athttp://research.cs.wisc.edu/bullying/data.html
6. A collection of insulting words can be found in the website: http://www.noswearing.com/dictionary

Fig. 5. Word Cloud Visualization of the List of Words with Negative Affective.

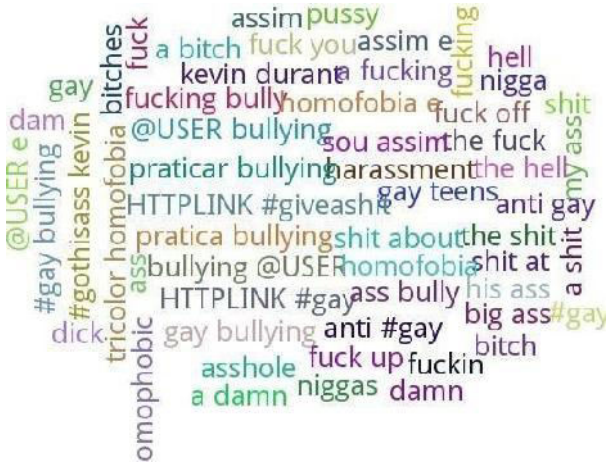


Fig. 6. Word Cloud Visualization of the Bullying Features in *Twitter* Datasets.

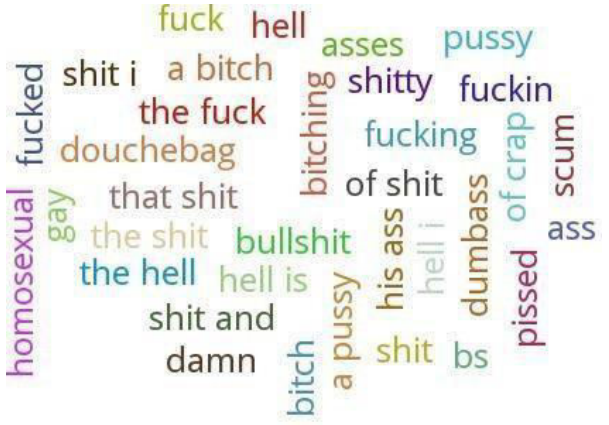


Fig. 7. Word Cloud Visualization of the Bullying Features in *MySpace* Datasets.

ISBN 978-93-88808-61-3

Experimental Results

In this section, we show a comparison of our proposed smSDA method with six benchmark approaches on *Twitter* and *MySpace* datasets. The average results, for these two datasets, on classification accuracy and F1 score are shown in Table 2. Figures 8 and 9 show the results of seven compared approaches on all *sub*-datasets constructed from *Twitter* and *MySpace* datasets, respectively. Since BWM does not require training documents, its results over the whole corpus are reported in Table 2. It is clear that our approaches outperform the other approaches in these two *Twitter* and *MySpace* corpora.

The first observation is that semantic BoW model (sBow) performs slightly better than BoW. Based on BoW, sBoWjust arbitrarily scale the bullying features by a factor of 2. This means that semantic information can boost theperformance of cyberbullying detection. For a fair comparison, the bul- lying features used in our method and sBoW are unified to be the same. Our approaches, especially smSDA, gains a significant performance improvement compared to sBoW. This is because bullying features only account for a small portion of all features used. It is difficult to learn robust features for small training data by intensifying each bullying features' amplitude. Our approach aims to find the cor- relation between normal features and bullying features by reconstructing corrupted data so as to yield robust features. In addition, Bullying Word Matching (BWM), as a simple and intuitive method of using semantic information, gives the worst performance. In BWM, the existence of bullying words are defined as rules for classification. It shows that only an elaborated utilization of such bullying words in- stead of a simple one can help cyberbullying detection.

We also compare our methods with two stat-of-arts text representation learning methods LSA and LDA. These two methods do not produce good performance on all datasets. This may be because that both methods belong to dimen- sionality reduction techniques, which are performed on the document-word occurrence matrix. Although the two meth- ods try to minimize the reconstruction error as our approach does, the optimization in LSA and LDA is conducted after dimensionality reduction. The reduced dimension is a key parameter to determine the quality of learned feature space. Here, we fix the dimension of latent space to 100. Therefore, a deliberate searching for this parameter which may im- prove the performances of LSA and LDA and the selection of hyperparameter itself is another tough research topic. Another reason may be that the data samples are small (less than 2000) and the length of each Internet message is short (For *Twitter*, maximum length is 140 characters),and thus the constructed document-word occurrence matrix may not represent the true co-occurrence of terms.

Deep learning methods including mSDA and smSDA generally outperform other standard approaches. This trend is particularly prominent in F1 measure because cyberbul- lying detection problems are class-imbalance. The larger improvements on F1 score verify the performance of our approach further. Deep learning models have achieved re- markable performance in various scenarios with its own robust feature learning ability [22]. mSDA is able to cap- ture the correlation between input features and combine

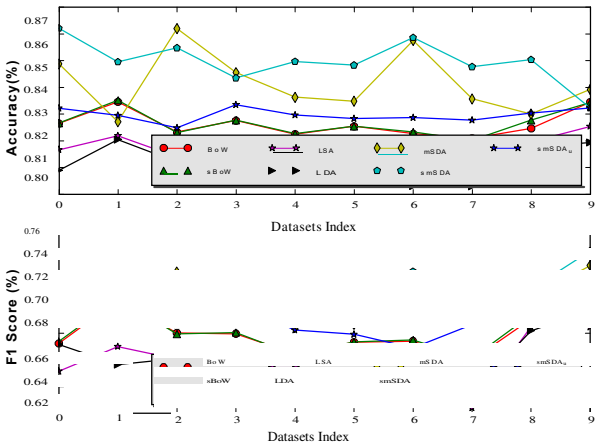


Fig. 8. Classification Accuracies and F1 Scores of All Compared Meth- ods on *Twitter* Datasets.

the correlated features by reconstructing masking feature values from uncorrupted feature values. Further, the stack- ing structure and the nonlinearity contribute to mSDA's ability for discovering complex factors behind data. Based on mSDA, our proposed smSDA utilizes semantic dropout noise and sparsity constraints on mapping matrix, in which the efficiency of training can be kept. This extension leads to a stable performance improvement on cyberbullying de- tection and the detailed analysis has been provided in the following section.

We compare the performances of smSDA and $smSDA_u$, which adopt biased semantic dropout noise and unbi- ased semantic dropout noise, respectively. The results have shown that $smSDA_u$ performs slightly worse than smSDA. This may be explained by the fact that the unbiased semantic dropout noise cancels the enhancement of bullying features. As shown in Eq. (14), the off-diagonal elements in the matrix $\mathbf{x}_i\tilde{\mathbf{x}}^T$ that are used to compute mapping weights are the same, which can not contribute to the reinforcement of bullying features.

Analysis of Semantic Extension

As shown in the section 4.3, the semantic extension can boost the performance on classification results for cyber- bullying detection. In this section, we discuss the advan- tages of this extension qualitatively. In our proposed smS- DA, because of the semantic dropout noise and sparsity constraints, the learned representation is able to discover the correlation between words containing latent bullying semantics. Table 3 shows the reconstruction terms of three example bullying words for mSDA and smSDA, respec- tively. In this example, one-hot vector is used as input, which represents a document containing one bullyingword. Table 3 lists the reconstructed terms in decreasing order of their feature values, which represents the strength of their correlations with the input word. The results are obtained using one layer architecture without non-linear activation considering the raw terms directly correspondto

ISBN 978-93-88808-61-3

TABLE 2
Accuracies (%), and F1 Scores (%) for Compared Methods on *Twitter* and *MySpace* Datasets. The Mean Values are Given, respectively. Bold Face Indicates Best Performance.

Dataset	Measures	BWM	BoW	sBow	LSA	LDA	mSDA	$smSDA_u$	smSDA
Twitter	Accuracies	69.3	82.6	82.7	81.6	81.1	84.1	82.9	**84.9**
	F1 Scores	16.1	68.1	68.3	65.8	66.1	70.4	69.3	**71.9**
MySpace	Accuracies	34.2	80.1	80.1	77.7	77.8	87.8	88.0	**89.7**
	F1 Scores	36.4	41.2	42.5	45.0	43.1	76.1	76.0	**77.6**

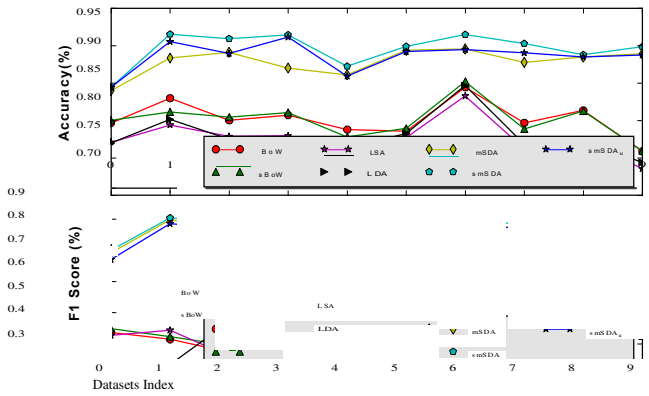


Fig. 9. Classification Accuracies and F1 Scores of All Compared Meth- ods on *MySpace* Datasets.

each output dimension under such a setting. It is shown that these reconstructed words discovered by smSDA are more correlated to bullying words than those by mSDA. For example, *fucking* is reconstructed by *because*, *friend*, *off*, *gets* in mSDA. Except *off*, the other three words seem to be unreasonable. However, in smSDA, *fucking* is reconstructed by *off*, *pissed*, *shit* and *of*. The occurrence of the term *of* may be due to the frequent misspelling in Internet writing. It is obvious that the correlation discovered by smSDA is more meaningful. This indicates that smSDA can learn the
words' correlations which may be the signs of bullying semantics, and therefore the learned robust features boost

the performance on cyberbullying detection.

5 Conclusion

This paper addresses the text-based cyberbullying detection problem, where robust and discriminative representations of messages are critical for an effective detection system.
By designing semantic dropout noise and enforcing spar- sity, we have developed semantic-enhanced marginalized denoising autoencoder as a specialized representation learn-
ing model for cyberbullying detection. In addition, word embeddings have been used to automatically expand and refine bullying word lists that is initialized by domain knowledge. The performance of our approaches has been experimentally verified through two cyberbullying corpora from social medias: *Twitter* and *MySpace*. As a next step we are planning to further improve the robustness of the

TABLE 3
Term Reconstruction on *Twitter* datasets. Each Row Shows Specific Bullying Word, along with Top-4 Reconstructed Words (ranked with their frequency values from top to bottom) via mSDA (left column) and smSDA (right column).

Bullying Words	Reconstructed Words for mSDA	Reconstructed Words for smSDA
bitch	*@USER shut friend tell*	*@USER HTTPLINK fuck up shut*
fucking	*because friend off gets*	*off pissed shit of*
shit	*some big with lol*	*abuse this shit shit lol big*

learned representation by considering word order in mes- sages.

Acknowledgments

We thank Junming Xu and Prof. Jerry Zhu for their *Twitter* datasets.

References

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.
[2] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta- analysis of cyberbullying research among youth." 2014.
[3] M. Ybarra, "Trends in technology-based sexual and non-sexual aggression over time and linkages to nontechnology aggression," *National Summit on Interpersonal Violence and Abuse Acrossthe Lifespan: Forging a Shared Agenda*, 2010.
[4] B. K. Biggs, J. M. Nelson, and M. L. Sampilo, "Peer relations in the anxiety–depression link: Test of a mediation model," *Anxiety, Stress, & Coping*, vol. 23, no. 4, pp. 431–447, 2010.
[5] S. R. Jimerson, S. M. Swearer, and D. L. Espelage, *Handbook of bullying in schools: An international perspective*. Routledge/Taylor & Francis Group, 2010.
[6] G. Gini and T. Pozzoli, "Association between bullying and psy- chosomatic problems: A meta-analysis," *Pediatrics*, vol. 123, no. 3, pp. 1059–1065, 2009.
[7] A. Kontostathis, L. Edwards, and A. Leatherman, "Text mining and cybercrime," *Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK*, 2010.
[8] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Compu- tational Linguistics, 2012, pp. 656–666.
[9] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proceedings of the 3rd Inter- national Workshop on Socially-Aware Multimedia*. ACM, 2014, pp. 3– 6.

[10] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis,and L. Edwards, "Detection of harassment on web 2.0," *Proceedings of the Content Analysis in the WEB*, vol. 2, pp. 1–7, 2009.

[11] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detec- tion of textual cyberbullying." in *The Social Mobile Web*, 2011.

[12] V. Nahar, X. Li, and C. Pang, "An effective approach for cy- berbullying detection," *Communications in Information Science and Management Engineering*, 2012.

[13] M. Dadvar, F. de Jong, R. Ordelman, and R. Trieschnigg, "Im- proved cyberbullying detection using gender information," in *Proceedings of the 12th -Dutch-Belgian Information Retrieval Workshop (DIR2012)*. Ghent, Belgium: ACM, 2012.

[14] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Im- proving cyberbullying detection with user context," in *Advances in Information Retrieval*. Springer, 2013, pp. 693–696.

[15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[16] P. Baldi, "Autoencoders, unsupervised learning, and deep archi- tectures," *Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7*, p. 43, 2012.

[17] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized de- noising autoencoders for domain adaptation," *arXiv preprint arX- iv:1206.4683*, 2012.

[18] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[19] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceed- ings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[21] T. Hofmann, "Unsupervised learning by probabilistic latent se- mantic analysis," *Machine learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[22] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[23] B. L. McLaughlin, A. A. Braga, C. V. Petrie, M. H. Moore *et al.*, *Deadly Lessons:: Understanding Lethal School Violence*. National Academies Press, 2002.

[24] J. Juvonen and E. F. Gross, "Extending the school ground- s?bullying experiences in cyberspace," *Journal of School health*, vol. 78, no. 9, pp. 496–505, 2008.

[25] M. Fekkes, F. I. Pijpers, A. M. Fredriks, T. Vogels, and S. P. Verloove-Vanhorick, "Do bullied children get ill, or do ill chil- dren get bullied? a prospective cohort study on the relationship between bullying and health-related symptoms," *Pediatrics*, vol. 117, no. 5, pp. 1568–1574, 2006.

[26] M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, and K. Araki, "Brute force works best against bullying," in *Proceedings of IJCAI 2015 Joint Workshop on Constraints and Preferences for Configuration and Recommendation and Intelligent Techniques for Web Personaliza- tion*. ACM, 2015.

[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[28] C. C. Paige and M. A. Saunders, "Lsqr: An algorithm for sparse linear equations and sparse least squares," *ACM Transactions on Mathematical Software (TOMS)*, vol. 8, no. 1, pp. 43–71, 1982.

[29] M. A. Saunders *et al.*, "Cholesky-based methods for sparse least squares: The benefits of regularization," *Linear and Nonlinear Con- jugate Gradient-Related Methods*, pp. 92–100, 1996.

[30] J. Fan and R. Li, "Variable selection via nonconcave penalized like- lihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[31] C. Vogel, *Computational Methods for Inverse Problems*. Society for Industrial and Applied Mathematics, 2002. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/1.9780898717570

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient esti- mation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their com- positionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[34] F. Godin, B. Vandersmissen, W. De Neve, and R. Van de Walle, "Named entity recognition for twitter microposts using distributed word representations," in *Proceedings of the Workshop on Noisy User-generated Text*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 146–153. [Online]. Available: http://www.aclweb.org/anthology/W15-4322

[35] T. H. Dat and C. Guan, "Feature selection based on fisher ratio and mutual information analyses for robust brain computer interface," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1. IEEE, 2007, pp. I–337.

[36] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[37] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[38] J. Sui, "Understanding and fighting bullying with machine learn- ing," Ph.D. dissertation, THE UNIVERSITY OF WISCONSIN- MADISON, 2015.

[39] J. Bayzick, A. Kontostathis, and L. Edwards, "Detecting the pres- ence of cyberbullying using computer software," in *Proceedings of the ACM WebSci'11*. Koblenz, Germany: ACM, June 2011, pp. 1–2.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Efficient and Expressive Keyword Search Over Encrypted Data in Cloud

L.Ananth Lakshmi
M.Tech.
Schholar, CSE Department,
Malla Reddy College of Engineering

Abstract—Searchable encryption allows a cloud server to conduct keyword search over encrypted data on behalf of the data users without learning the underlying plaintexts. However, most existing searchable encryption schemes only support single or conjunctive keyword search, while a few other schemes that are able to perform expressive keyword search are computationally inefficient since they are built from bilinear pairings over the composite-order groups. In this paper, we propose an expressive public-key searchable encryption scheme in the prime-order groups, which allows keyword search policies (i.e., predicates, access structures) to be expressed in conjunctive, disjunctive or any monotonic Boolean formulas and achieves significant performance improvement over existing schemes. We formally define its security, and prove that it is selectively secure in the standard model. Also, we implement the proposed scheme using a rapid prototyping tool called Charm [37], and conduct several experiments to evaluate it perform ance. The results demonstrate that our scheme is much more efficient than the ones built over the composite-order groups.

Index Terms—Searchable encryption, cloud computing, expressiveness, attribute-based encryption.

1 INTRODUCTION

Consider a cloud-based healthcare information system that hosts outsourced personal health records (PHRs) from various healthcare providers. The PHRs are encrypted in order to comply with privacy regulations like HIPAA. In order to facilitate data use and sharing, it is highly desir- able to have a searchable encryption (SE) scheme which allows the cloud service provider to search over encrypted PHRs on behalf of the authorized users (such as medical researchers or doctors) without learning information about the underlying plaintext. Note that the context we are con- sidering supports private data sharing among multiple data providers and multiple data users. Therefore, SE schemes in the private-key setting [1], [2], [3], which assume that a single user who searches and retrieves his/her own data, are not suitable. On the other hand, private information retrieval (PIR) protocols [4], [5], [6], which allow users to retrieve a certain data-item from a database which publicly stores data without revealing the data-item to the database administrator, are also not suitable, since they require the data to be publicly available. In order to tackle the keyword search problem in the cloud-based healthcare information system scenario, we resort to public-key encryption with keyword search (PEKS) schemes, which is firstly proposed in [7]. In a PEKS scheme, a ciphertext of the keywords called "PEKS ciphertext" is appended to an encrypted PHR. To retrieve all the encrypted PHRs containing a keyword, say "Diabetes", a user sends a "trapdoor" associated with a search query on the keyword "Diabetes" to the cloud service provider, which selects all the encrypted PHRs containing the keyword "Diabetes" and returns them to the user while without learning the underlying PHRs. However, the solu- tion in [7] as well as other existing PEKS schemes which improve on [7] only support equality queries [8].

Set intersection and meta keywords[1] [9], [10] can be used for conjunctive keyword search. However, the approach based on set intersection leaks extra information to the cloud server beyond the results of the conjunctive query, whilst the approach using meta keywords require 2^m meta keywords to accommodate all the possible conjunctive queries for m keywords. In order to address the above deficiencies in conjunctive keyword search, schemes such as the ones in [11], [12] were put forward in the public-key setting.

Ideally, in the practical applications, search predicates (i.e., policies) should be expressive such that they can be expressed as conjunction, disjunction or any Boolean for- mulas[2] of keywords. In the above cloud-based healthcare system, to find the relationship between diabetes and age or weight, a medical researcher may issue a search query with an access structure (i.e., predicate) ("Illness = Diabetes" AND ("Age = 30" OR "Weight = 150-200")). SE schemes supporting expressive keyword access structures were p- resented in [8], [13], [14], [15]. Unfortunately, the scheme in [13] has exponentially increasing complexity [16], while the schemes in [8], [14], [15] are based on the inefficient bilinear pairing over composite-order groups [17]. Though there exist techniques [17] to convert pairing-based schemes from composite-order groups to prime-order groups, there is still a significant performance degradation due to the

1. Meta keywords are composed of several keywords. For example, a that contains the document keywords "Bob", "urgent" and "finance" may be augmented with the meta-keyword "Bob: urgent: finance"

required size of the special vectors [18].

In this paper, we propose a public-key based expressive SE scheme in prime-order groups, which is especially suit- able for keyword search over encrypted data in scenarios of multiple data owners and multiple data users such as the cloud-based healthcare information system that hosts outsourced PHRs from various healthcare providers.

Overview of Our Proposed Scheme

Our expressive SE scheme consists of a trusted trapdoor generation center which publishes a public system parame- ter and keeps a master key in secret, a cloud server which stores and searches encrypted data on behalf of data users, multiple data owners who upload encrypted data to the cloud, and multiple data users who would like to retreive encrypted data containing certain keywords. To outsource an encrypted document to the cloud, a data owner appends the encrypted document with keywords encrypted under the public parameter and uploads the combined encrypted document and encrypted keywords to the cloud. To retrieve all the encrypted documents containing keywords satisfying a certain access structure (i.e., predicate or policy) such as ("Illness = Diabetes" AND ("Age = 30" OR "Weight = 150-200")), a data user first obtains a trapdoor associated with the access structure from the trapdoor generation center and then sends the trapdoor to the cloud server. The latter will conduct the search and return the corresponding encrypted documents to the data user.

The basic idea of our scheme is to modify a key-policy attributed-based encryption (KP-ABE) scheme constructed from bilinear pairing over prime-order groups. Without loss of generality, we will use the large universe KP-ABE scheme selectively secure in the standard model proposed by Rouse- lakis and Waters in [18] to illustrate our construction during the rest of the paper. In KP-ABE, a ciphertext is computed with respect to a set of attributes and an access policy is encoded into a user's private key. A ciphertext can be decrypted by a private key only if the set of attributes associ- ated with the ciphertext satisfies the access policy associated with the private key. Access policies in [18] can be very expressive, supporting any monotonic Boolean formulas. At first sight, a KP-ABE scheme can be transformed to an expressive SE scheme by treating attributes as keywords to be searched, by directly transforming the key generation algorithm on attribute access structures to a trapdoor gener- ation algorithm on keyword search predicates, and by using the decryption algorithm to test whether keywords in a ciphertext satisfy the predicate in a trapdoor. However, KP- ABE schemes (e.g., [18], [19]) are not designed to preserve privacy of attributes (keywords) associated with ciphertexts. Specifically, given the public parameter and a ciphertext, the attributes (keywords) in the ciphertext can be discerned by anyone. In the following, to keep our description compact and consistent, we will use access structure, policy and predicate interchangeably.

In order to hide keywords in a ciphertext, inspired by the "linear splitting" technique in [20], we firstly split ci- phertext components corresponding to every keyword into two randomized complementary components. Thus, even though the ciphertext still contains information about the keywords, this information is computationally infeasible to obtain from the public parameter and the ciphertext. We secondly re-randomize trapdoor components corresponding to every keyword associated with an access structure to match the splitted components in the ciphertext.

In addition to hiding keywords in ciphertexts, we also need to preserve keyword privacy in a trapdoor which contains an access structure as a component. First, to p- reserve keyword privacy in an access structure, we adopt the method in [21] to divide each keyword into a generic name and a keyword value. Since keyword values are much more sensitive than the generic keyword names, the key- word values in an access structure are not disclosed to the cloud server, whereas a partial hidden access structure with only generic keyword names is included in a trapdoor and sent to the cloud server. Take the aforementioned keyword access structure ("Illness = Diabetes" AND ("Age = 30" OR "Weight = 150-200")) as an instance, "Illness", "Age" and "Weight" are the generic names whilst "Diabetes", "30" and "200" are the keyword values. Consequently, the partial hidden access structure ("Illness" AND "Age" OR "Weight") is included in the trapdoor. Second, as in all the PEKS schemes, trapdoors are subject to the offline keyword dictionary guessing attacks. That is, anyone who knows a trapdoor and the public parameter may discover the keyword values embedded in the trapdoor by launching exhaustive searching attacks on keyword values. As a rem- edy to such attacks, we assign a designated cloud server as introduced in [22] to perform the searching operations. We equip this designated server with a public and private key pair of which the public key will be used in trapdoor generation such that it is computationally infeasible for anyone without knowledge of the privacy key to derive keywords information from the trapdoor. Thus, trapdoors can be delivered to the cloud server over a public channel.

We define a security model for expressive SE, which takes into account all adversarial capabilities of the standard SE security notion. The adversary is able to learn trapdoors over access structures of its choice, but it should not be able to learn any information about the keyword values in the challenge ciphertext. Note that since the Rouselakis-Waters KP-ABE scheme [18], which the proposed SE scheme isbuilt upon, is selectively secure, our expressive SE scheme can only be proved to be selectively secure where the adversary has to commit the challenge keyword set in advance.

Contributions

Below we briefly summarize our contributions in this paper.

- We propose the first expressive SE scheme in the public-key setting from bilinear pairings in *prime- order* groups. As such, our scheme is not only capable of expressive multi-keyword search, but also signif- icantly more efficient than existing schemes built in composite-order groups.
- Using a randomness splitting technique, our scheme achieves security against offline keyword dictionary guessing attacks to the ciphertexts. Moreover, to preserve the privacy of keywords against offline keyword dictionary guessing attacks to trapdoors, we divide each keyword into keyword name and

ISBN 978-93-88808-61-3

keyword value and assign a designated cloud server to conduct search operations in our construction.

- We formalize the security definition of expressive SE, and formally prove that our proposed expressive SE scheme is selectively secure in the standard model.
- We implement our scheme using a rapidly proto- typing tool called Charm, and conduct extensive experiments to evaluate its performance. Our results confirm that the proposed scheme is sufficiently effi- cient to be applied in practice.

Related Work

Public-Key Encryption with Keyword Search. After Boneh et al. [7] initiated the study of public-key encryption with keyword search (PEKS), several PEKS constructions were put forth using different techniques or considering differ- ent situations [8], [11], [12], [13], [14], [15], [22], [23], [24], [25], [26], [27], [28], [29]. They aim to solve two cruces in PEKS: (1) how to make PEKS secure against offline keyword dictionary guessing attacks; and (2) how to achieve expressive searching predicates in PEKS. In terms of the offline keyword dictionary guessing attacks, which requires that no adversary (including the cloud searching server) can learn keywords from a given trapdoor, to the best of our knowledge, such a security notion is very hard to be achieved in the public-key setting [30]. Regarding the expressive search, there are only few works in PEKS [8], [13], [14], [15]. Unfortunately, the construction in [13] is built on the basis of inner-product predicate encryption [16], and the constructions in [8], [14], [15] are built from the pairings in composite-order group. Therefore, they are not sufficiently efficient to be adopted in the practical world [16], [17].

Moreover, the number of keywords allowed in these search- able schemes are predefined in the system setup phase. We compare our scheme to other keyword search schemes in Table 1. It is straightforward to see that compared to the existing ones, our construction make a good balance in that it allows unbounded keywords, supports expressive access structures, and is built in the prime-order groups.

Private-key Searchable Encryption. In a private-key SE setting, a user uploads its private data to a remote database and keeps the data private from the remote database ad- ministrator. Private-key SE allows the user to retrieve all the records containing a particular keyword from the re- mote database [1], [2], [3]. However, as the name suggests, private-key SE solutions only apply to scenarios where data owners and data users totally trusted each other.

Private Information Retrieval. With respect to public database such as stock quotes, where the user is unaware of it and wishes to search for some data-item without revealing to the database administrator which item it is, private infor- mation retrieval (PIR) [4], [5], [6] protocols were introduced, which allow a user to retrieve data from a public database with far smaller communication then just downloading the entire database. Nevertheless, in our context, the database is not publicly available, the data is not public, so the PIR solutions cannot be applied.

Organization

The remainder of this paper is organized as follows. In Sec- tion 2, we briefly review some of the notions and definitions to be used in the paper. In Section 3, after depicting the sys- tem architecture for our expressive keyword search system, we give a concrete expressive keyword search scheme. In Section 4, we discuss the properties and several extensions of our expressive keyword search scheme. We implement our scheme and compare it with related works in Section 5. We conclude the paper in Section 6.

2 PRELIMINARIES

In this section, we review some basic cryptographic notions and definitions that are to be used later.

Bilinear Pairings and Complexity Assumptions

Let tt be a group of prime order p with a generator g. Let $\hat{e}$: $tt \times tt \mapsto$ be an efficiently computable bilinear pairing function satisfying the following properties [31].

- Bilinear: for all $g \in tt$, and $a, b \in Z_p^*$, we have $\hat{e}(g^a, g^b) = \hat{e}(g, g)^{ab}$.
- Non-degenerate: $\hat{e}(g, g) \neq 1$.

Decisional Bilinear Diffie-Hellman Assumption [31]. The decisional Bilinear Diffie-Hellman (BDH) problem is that for any probabilistic polynomial-time algorithm, given g, g^a, g^b, g^c, it is difficult to distinguish $(g, g^a, g^b, g^c, \hat{e}(g, g)^{abc})$ from (g, g^a, g^b, g^c, Z), where $g \in tt, Z \in tt$, $a, b, c \in Z_p^*$ chosen independently and uniformly at random.

Decisional $(q-2)$ Assumption [18]. Let q be an integer. The decisional $(q-2)$ problem is that for any probabilistic polynomial-time algorithm, given $\vec{A} = g, g^x, g^y, g^z, g^{(xz)^2}$, $g^{b_i}, g^{xzb_i}, g^{xz/b_i}, g^{x^2zb_i}, g^{y/b^2}, g^{y^2/b^2}$ $\forall i \in [q]$, $g^{xzb_i/b_j}, g^{yb_i/b^2}, g^{xyzb_i/b_j}, g^{(xz)^2b_i/b_j}$ $\forall i, j \in [q], i \neq j$, it is difficult to distinguish $(\vec{A}, \hat{e}(g, g)^{xyz})$ from $(\vec{A}, Z)$, where $g \in tt, Z \in tt, x, y, z, b_1, ..., b_q \in Z_p^*$ chosen independently and uniformly at random.

Decisional Linear Assumption [32]. The decisional lin- ear problem is that for any probabilistic polynomial-time algorithm, given $g, g^{x_1}, g^{x_2}, g^{x_1x_3}, g^{x_2x_4}$, it is difficult to distinguish $(g, g^{x_1}, g^{x_2}, g^{x_1x_3}, g^{x_2x_4}, g^{x_3+x_4})$ from $(g, g^{x_1}, g^{x_2}, g^{x_1x_3}, g^{x_2x_4}, Z)$, where $g, Z \in tt, x_1, x_2, x_3, x_4 \in Z_p^*$ chosen independently and uniformly at random.

Access Structures and Linear Secret Sharing

Following the definition in [33], [34], we describe the notions of access structures and linear secret sharing schemes.

Definition 1. (Access Structure). Let $\{P_1, ..., P_n\}$ be a set of parties. A collection $A \subseteq 2^{\{P_1,...,P_n\}}$ is monotone if $\forall B, C$: if $B \in A$ and $B \subseteq C$, then $C \in A$. An (monotone) access structure is a (monotone) collection A of non-empty subsets of $\{P_1, ..., P_n\}$, i.e., $A \subseteq 2^{\{P_1,...,P_n\}} \setminus \{\emptyset\}$. The sets in A are called the authorized sets, and the sets not in A are called the unauthorized sets.

In our construction, we only consider monotone access structures. Notice that general access structures in large uni- verse ABE can be realized by splitting the attribute universe in half and treating the attributes of one half as the negated

ISBN 978-93-88808-61-3

TABLE 1
Comparisons of expressive keyword search schemes.

	Keyword Privacy	Expressiveness	Bilinear Group	Security	Unbounded keywords
BCOP04 [7]	keyword guessing attacks on trapdoors	AND	prime	full random oracle	yes
KSW13 [16]	keyword guessing attacks on trapdoors	AND, OR	composite	full standard model	no
LZDLC13 [8]	keyword guessing attacks on trapdoors	AND, OR	composite	full standard model	no
LHZF14 [14]	no keyword guessing attacks on trapdoors	AND, OR, NOT	composite	full standard model	no
Our scheme	keyword guessing attacks on trapdoors by designated server only	AND, OR	prime	selective standard model	yes

(NOT) versions of the attributes in the other half [35]. Also, it has been presented in [36], [37] how to describe non- monotonic access structures in terms of monotonic access structures with negative (NOT) shares.

Definition 2. (Linear Secret Sharing Schemes). Let P be a set of parties. Let M be a matrix of size $l \times$. Let $\rho : 1, \{ ..., l \}$ $\rightarrow P$ be a function that maps a row to a party for labeling. A secret sharing scheme Π over a set of parties P is a linear secret-sharing scheme (LSSS) over Z_p if

1) The shares for each party form a vector over Z_p.
2) There exists a matrix M which has l rows and n columns called the share-generating matrix for Π. For $i = 1, ..., l$, the x-th row of matrix M is labeled

by a party $\rho(i)$, where $\rho : \{1, ..., l\} \rightarrow P$ is a function that maps a row to a party for labeling. Considering that the column vector $\overrightarrow{v} = (\mu, r_2, ..., r_n)$, where $\mu \in Z_p$ is the secret to be shared and $r_2, ..., r_n \in Z_p$ are randomly chosen, then $M\overrightarrow{v}$ is the vector of l shares of the secret μ according to Π. The share $(M\overrightarrow{v})_i$ belongs to party $\rho(i)$.

It has been noted in [33] that every LSSS also enjoys the linear reconstruction property. Suppose that $\Pi$is an LSSS for an access structure A. Let **A** be an authorized set, anddefine $I \subseteq \{1, ..., l\}$ as $I = \{i | \rho(i) \in \mathbf{A}\}$. Then the vector $(1, 0, ..., 0)$ is in the span of rows of M indexed by I, and there exist constants $\{w_i \in Z_p\}_{i \in I}$ such that, for any valid shares $\{v_i\}$ of a secret μ according to Π, we have $\sum_{i \in I} w_i v_i = \mu$. These constants $\{w_i\}$ can be found in the polynomial time in terms of the size of the share-generating matrix M [38]. On the other hand, for an unauthorized set $\mathbf{A}^J$, no such constants $\{w_J\}$ exist. Also, in this case it is true that if $I^J = \{i | \rho(i) \in \mathbf{A}^J\}$, there exists a vector w such that its first component w is any non zero element in Z_p and $< M_i, w > = 0$ for all $i \in I^J$, where M_i is the i-th row of M [18].

Boolean Formulas [33]. Access structures can also be described in terms of monotonic boolean formulas. LSSS access structures are more general, and can be derived from representations as boolean formulas. There are techniques to convert any monotonic boolean formula into a correspond- ing LSSS matrix[3]. The boolean formula can be represented as an access tree, where the interior nodes are AND and OR gates, and the leaf nodes denote attributes. The number of the rows in the corresponding LSSS matrix will be the same as the number of the leaf nodes in the access tree.

3. We give an example on how to convert a boolean formula into an equivalent LSSS matrix in Appendix C.

3 Efficient and Expressive Keyword Search with Unbounded Keywords

In this section, we describe the system model, design goals, threat model and algorithms of our expressive SE scheme.

System Model and Design Goals

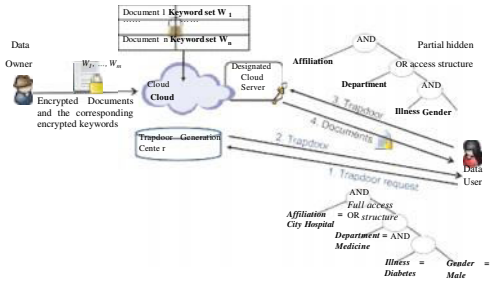


Fig. 1. Architecture of expressive keyword search system.

The architecture of our keyword search system is shown in Fig. 1, which is composed of four entities: a trusted trapdoor generation centre who publishes the system pa- rameter and holds a master private key and is responsible for trapdoor generation for the system, data owners who outsource encrypted data to a public cloud, data users who are privileged to search and access encrypted data, and a designated cloud server who executes the keyword search operations for data users. To enable the cloud server to search over ciphertexts, the data owners append every encrypted document with encrypted keywords[4]. A data user issues a trapdoor request by sending a keyword access structure to the trapdoor generation centre which generates and returns a trapdoor corresponding to the access struc- ture. We assume that the trapdoor generation centre has a separate authentication mechanism to verify each data user and then issue them the corresponding trapdoors. After

4. Note that each keyword is composed of a generic name and a keyword value.

obtaining a trapdoor, the data user sends the trapdoor and the corresponding partial hidden access structure (i.e., the access structure without keyword values) to the designated cloud server. The latter performs the testing operations between each ciphertext and the trapdoor using its private key, and forwards the matching ciphertexts to the data user. As mentioned earlier, a ciphertext created by a data owner consists of two parts: the encrypted document generated using an encryption scheme and the encrypted keywords generated using our SE scheme. From now on, we only consider the latter part of the encrypted document, and ignore the first part since it is out of the scope of this paper.

In summary, the design goals of our expressive SE scheme are fourfold.

- **Expressiveness.** The proposed scheme should sup- port keyword access structures expressed in any Boolean formula with AND and OR gates.
- **Efficiency.** The proposed scheme should be ade- quately efficient in terms of computation, commu- nication and storage for practical applications.
- **Keyword privacy.** First, a ciphertext without its corresponding trapdoors should not disclose any information about the keyword values it contains to the cloud server and outsiders. Second, a trapdoor should not leak information on keyword values to any outside attackers without the private key of the designated cloud server. We capture this notion of security for the SE scheme in terms of semantic security to ensure that encrypted data does not reveal any information about the keyword values, which we call "selective indistinguishability against chosen keyword-set attack (selective IND-CKA security)" (See Appendix A).
- **Provable security.** The security of the proposed scheme should be formally proved under the stan- dard model rather than the informal analysis.

Threat Model

We assume that the trapdoor generation centre is a trusted entity.

The cloud server is assumed to be "honest-but-curious", i.e., it will honestly follow the protocol but it is also curious to learn any private information from the data stored in the cloud. Data owners are assumed to honestly store their data, while data users are not trusted, and they can even collude with a malignant cloud server in order to discover private information of other parties. We assume that the trusted trapdoor generation centre is equipped with a separate authentication mechanism to verify data users before issuing trapdoors to users. Also, we assume that all adversaries have bounded computational capability, so they cannot break the aforementioned difficult problems.

Construction

In the system, the trusted trapdoor generation centre is given a public parameter and a master private key generated by the Setup algorithm, and uses the Trapdoor algorithm to generate a trapdoor T_{M} for some keyword set associated with an access structure $(\mathrm{M}, \rho, \{W_{\rho(i)}\})$ at the request of a privileged data user, where M is an access matrix, ρ is the function that associates the rows of M to the generic names of keywords, and $\{W_{\rho(i)}\}$ are the corresponding keyword values[5]. The cloud server is given a public and private key pair created by the sKeyGen algorithm, and will input the trapdoor given by a data user and its private key to the Test algorithm to determine whether a document contains the keywords satisfying the keyword access structure $(\mathrm{M}, \rho, \{W_{\rho(i)}\})$ specified by the data user.

Let tt be a group of prime order p with a generator g, and $\hat{e}: tt \times tt \rightarrow tt_1$ be the bilinear map. On the basis of the KP-ABE scheme proposed by Rouselakis and Waters [18], which we will refer to as the Rouselakis-Waters KP- ABE scheme, we describe our expressive and unbounded SE system in the prime-order groups as follows.

- Setup. This algorithm takes the security parameter 1^λ as input. It randomly chooses a group tt of prime order p, a generator g and random group elements $u, h, w \in tt$. Also, it randomly chooses $\alpha, d_1, d_2, d_3, d_4 \in Z_p$, and computes $g_1 = g^{d_1}$, $g_2 = g^{d_2}$, $g_3 = g^{d_3}$, $g_4 = g^{d_4}$. Finally, it publishes the public parameter $pars = (H, g, u, h, w, g_1, g_2, g_3, g_4, \hat{e}(g, g)^\alpha)$, where H is a collision-resistant hash function that maps elements in tt_1 to elements in tt, and keeps the master private key $msk = (\alpha, d_1, d_2, d_3, d_4)$.
- sKeyGen. This algorithm takes the public parameter $pars$ as input. It randomly chooses $\gamma \in Z_p^*$, and outputs the public and private key pair $(pk_s, sk_s) = (g^\gamma, \gamma)$ for the server.
- Trapdoor. This algorithm takes the public param- eter $pars$, the server public key pk_s, the master private key msk and an LSSS access structure $(\mathrm{M}, \rho, \{W_{\rho(i)}\})$[6] as input, where M is an $l \times n$ matrix over Z_p, the function ρ associates the rows of M to generic keyword names, and $\{W_{\rho(i)}\}$ are the corre- sponding keyword values. Let M_i be the i-th row of M for $i \in \{1, \ldots, l\}$ and $\rho(i)$ be the keyword name associated with this row by the mapping ρ. It randomly chooses a vector $\vec{y} = (\alpha, y_2, \ldots, y_n)^\perp$ where $y_2, \ldots, y_n \in Z_p$, $r, r' \in Z_p$, $t_{1,1}, t_{1,2}, \ldots, t_{l,1}, t_{l,2} \in Z_p$, computes $T = g^r$, $T' = g^{r'}$, and outputs the trapdoor $T_{\mathrm{M},\rho} = (\mathrm{M}, \rho), T, T', \{T_{i,1}, T_{i,2}, T_{i,3}, T_{i,4}, T_{i,5}, T_{i,6}\}_{i \in [1,l]}$ as
$T_{i,1} = g^{v_i} w^{d_1 d_2 t_{i,1} + d_3 d_4 t_{i,2}}$,
$T_{i,2} = H(\hat{e}(pk_s, T')^r) \cdot g^{d_1 d_2 t_{i,1} + d_3 d_4 t_{i,2}}$,
$T_{i,3} = ((u^{W_{\rho(i)}} h)^{t_{i,1}})^{-d_2}$, $T_{i,4} = ((u^{W_{\rho(i)}} h)^{t_{i,1}})^{-d_1}$,
$T_{i,5} = ((u^{W_{\rho(i)}} h)^{t_{i,2}})^{-d_4}$, $T_{i,6} = ((u^{W_{\rho(i)}} h)^{t_{i,2}})^{-d_3}$,
where $v_i = \mathrm{M}_i \cdot \vec{y}$ is the share associated with the row M_i of the access matrix M. Note that only (M, ρ) is included in the trapdoor $T_{\mathrm{M},\rho}$.
- Encrypt. This algorithm takes the public parameter $pars$ and a keyword set $\mathbf{W}$ (each keyword is denoted as $N_i = W_i$, where N_i is the generic keyword name and W_i is the corresponding keyword value) as in- put. Let m be the size of $\mathbf{W}$, and $W_1, \ldots, W_m \in Z_p$ be

5. To be distinguishable from the keyword value W_i in a ciphertxt, we use $W_{\rho(i)}$ to denote the keyword value in a trapdoor.

6. For the details about how to convert a boolean formula into an equivalent LSSS matrix, please refer to [33].

ISBN 978-93-88808-61-3

the values of **W**. It randomly chooses $\mu, s_{1,1}, s_{1,2}, \ldots, s_{m,1}, s_{m,2}, z_1, \ldots, z_m \in Z_p$, and outputs a ciphertext CT = $C, D, \{(D_i, E_{i,1}, E_{i,2}, F_{i,1}, F_{i,2})\}_{i\in[1,m]}$ as

$$C = \hat{e}(g, g)^{\alpha\mu}, \quad D = g^{\mu},$$
$$D_i = w^{-\mu}(u^{W_i}h)^{z_i}, \quad E_{i,1} = g_1^{z_i - s_{i,1}},$$
$$E_{i,2} = g_2^{s_{i,1}}, \quad F_{i,1} = g_3^{z_i - s_{i,2}}, \quad F_{i,2} = g_4^{s_{i,2}}.$$

Note that in the implementation, to efficiently conduct keyword search, the ciphertext will be stored along with the generic names $\{N_i\}$ corresponding to keyword values $\{W_i\}$. Thus, before performing the Test algorithm on the encrypted keyword values, the matching on the keyword names will be executed, thereby reducing the searching time.

- Test. This algorithm takes the public parameter *pars*, the server private key sk_S, a ciphertext $C, D, \{(D_i, E_{i,1}, E_{i,2}, F_{i,1}, F_{i,2})\}$ on a keyword set **W** and a trapdoor $T_{M,\rho}$ associated with an access structure (M, ρ, $\{W_{\rho(i)}\}$) as input. It calculates $I_{M,\rho}$ from (M, ρ), which is a set of minimum subsets satisfying (M, ρ). It then checks whether there is an $\mathbf{I} \in I_{M,\rho}$ satisfying

$$\prod_{i\in\mathbf{I}} \left(\hat{e}(D, T_{i,1})\hat{e}\left(D_i, \frac{T_{i,2}}{H(\hat{e}(T, T^{J})^{\gamma})}\right)\hat{e}(E_{i,1}, T_{i,3})\right.$$
$$\left.\hat{e}(E_{i,2}, T_{i,4})\hat{e}(F_{i,1}, T_{i,5})\hat{e}(F_{i,2}, T_{i,6})\right)^{\Sigma_{w_i}} = C,$$

where $\sum_{i\in\mathbf{I}} w_i \mathbf{M}_i = (1, 0, \ldots, 0)$. It outputs 0 if no element in $I_{M,\rho}$ satisfying this equation or 1 otherwise.

Remarks. In our construction, the term g^{z_i} in the original construction [18] is split into $g^{z_i - s_{i,1}}$ and $g^{s_{i,1}}$. Thus, W_i is subtly hidden from the ciphertext. To see this, we have

$$\hat{e}(D_i, g_1) = \hat{e}(w, D)^{-1} \cdot \hat{e}(u^{W_i}h, g^{r_i}),$$

where the term $g_1^{z_i}$ is embedded in $g_1^{z_i - s_{i,1}}$, which cannot be computed from $g_1^{z_i - s_{i,1}}$ without knowing the value of $g_1^{s_{i,1}}$. Nevertheless, given $g_2^{s_{i,1}}$, it is difficult to compute the value of $g_1^{s_{i,1}}$. Similarly, the result works with $g_3^{z_i - s_{i,2}}$, $g_4^{s_{i,2}}$ as well. Note that some redundant elements as g_3, g_4, $T_{i,5}$, $T_{i,6}$, $F_{i,1}$, $F_{i,2}$ are introduced in our scheme to make the proof go smoothly.

Correctness

If the keyword set **W** embedded in a ciphertext satisfies the access structure associated with the trapdoor, we will have $\sum_{i\in\mathbf{I}} v_i w_i = \alpha$. Therefore,

$$\prod_{i\in\mathbf{I}} \left(\hat{e}(D, T_{i,1})\hat{e}\left(D_i, \frac{T_{i,2}}{H(\hat{e}(T, T^{J})^{\gamma})}\right)\hat{e}(E_{i,1}, T_{i,3})\right.$$
$$\left.\hat{e}(E_{i,2}, T_{i,4})\hat{e}(F_{i,1}, T_{i,5})\hat{e}(F_{i,2}, T_{i,6})\right)^{\Sigma_{w_i}}$$
$$= \prod_{i\in\mathbf{I}} \hat{e}(g^{\mu}, g^{v_i w^{d_1 d_2 t_{i,1} + d_3 d_4 t_{i,2}}})^{w_i}$$
$$\cdot \hat{e}(w^{-\mu}(u^{W_i}h)^{z_i}, g^{d_1 d_2 t_{i,1} + d_3 d_4 t_{i,2}})^{w_i}$$
$$\cdot \hat{e}(g_1^{z_i - s_{i,1}}, ((u^{W_{\rho(i)}}h)^{t_{i,1}})^{-d^2})^{w_i}$$
$$\cdot \hat{e}(g_2^{s_{i,1}}, ((u^{W_{\rho(i)}}h)^{t_{i,1}})^{-d^1})^{w_i}$$
$$\cdot \hat{e}(g_3^{z_i - s_{i,2}}, ((u^{W_{\rho(i)}}h)^{t_{i,2}})^{-d^4})^{w_i}$$
$$\cdot \hat{e}(g_4^{s_{i,2}}, ((u^{W_{\rho(i)}}h)^{t_{i,2}})^{-d^3})^{w_i}$$
$$= \hat{e}(g, g)^{\mu \sum_{i\in\mathbf{I}} v_i w_i} = \hat{e}(g, g)^{\alpha\mu}$$

Security Proof

Theorem 1. Under the decisional BDH assumption, the $(q - 2)$ assumption and the decisional linear assumption, our scheme is selectively indistinguishable under chosen keyword-set attacks (selective IND-CKA security).

Proof. The details of the selective IND-CKA security definition and its proof are given in Appendix B. The proof is divided into two parts, depending on the role of the adversary. In the first part, the adversary is assumed to be an outside attacker, and in the second part, the adversary is assumed to be the cloud server who performs search operations.

4 Discussion and Analysis

In this section, we discuss the properties as well as extensions of our expressive SE scheme.

Keyword Privacy

Keyword Value Guessing Attacks on Ciphertexts. Below we briefly review the encryption algorithm of the KP-ABE scheme in [18], and then show that there exists a keyword value guessing attack if it is directly transformed into a searchable encryption scheme.

Encrypt. Let m denote the size of **W** and $W_1, \ldots, W_m \in Z_p$ be the specific values of **W**. It randomly chooses $\mu, z_1, \ldots, z_m \in Z_p$, and outputs a ciphertext CT = $C, D, \{(C_i, D_i)\}_{i\in[1,m]}$.

$$C = \hat{e}(g, g)^{\alpha\mu}, \quad D = g^{\mu},$$
$$\forall i \in [m] \quad C_i = w^{-\mu}(u^{W_i}h)^{z_i}, \quad D_i = g^{z_i},$$

where $g, u, h, w, \hat{e}(g, g)^{\alpha}$ are the public parameters.

Given a ciphertext CT = $C, D, \{(C_i, D_i)\}_{i\in[1,m]}$, an adversary can easily determine whether a keyword value W^{J} is incorporated in the ciphertext by checking whether the following equation holds.

$$\hat{e}(C_i, g) = \hat{e}(w^{-1}, D) \cdot \hat{e}(u^{W^{t}}h, D_i).$$

In order to prevent such attacks, in our construction, we use a "linear splitting" technique [20] on each keyword value related component of the ciphertext, and then re-randomize the components upon each keyword value in the trapdoor. The former step prevents keyword value guessing attacks to the ciphertext while the latter step allows the trapdoor to be used for testing keyword values in the ciphertext.

Keyword Value Guessing Attacks on Trapdoors. Concerning this security requirement, we need to tackle two problems in our construction. First, keywords associated with a trapdoor must be hidden from the access structure. We address this problem by separating each keyword into a generic name and a keyword value, i.e., each keyword is in the form of "generic name = keyword value", and a partial hidden access structure, i.e., the full access structure with keyword values being removed (See Fig. 1) is incorporated in a trapdoor and given to the designated cloud server. Second, the entire trapdoor should be immune to the offline keyword value guessing attacks [25]. In our SE system,

ISBN 978-93-88808-61-3

we resort to a weaker security notion by requiring that a trapdoor will not disclose information about the keyword values in the ciphertext to an adversary excluding the cloud server who executes the searching operations. We assign a designated cloud server [22] to conduct search and equip it with a public and private key pair. Since the components in a trapdoor are tied with the public key of the server, only the designated cloud server with the corresponding private key is capable to learn the keyword values hidden in the trapdoor by performing offline guessing attacks.

Unbounded keyword search

In "small universe" KP-ABE constructions [18], the size of the keyword space were polynomially bounded in the security parameter and the keywords were fixed at the setup phase. Moreover, the sizes of the public parameter- s grow linearly with the number of keywords [8], [14], [15]. On the contrary, in "large universe" constructions, the size of the keyword space can be exponentially large, so it is much more desirable in the real-world applications. Our construction of the expressive SE scheme inherits the advantages of the Rouselakis-Waters scheme [18]. Thus, it is straightforward to see that in our SE scheme, the size of the public parameter is immutable with the number of keywords, and the number of the keywords allowed for the system is unlimited and can be freely set.

Extensions

Our expressive SE system can be extended in several ways.

- Expressive searchable encryption for the range search. Range search is an important requiremen- t for searchable encryption in many applications. By defining keywords in a hierarchical manner as shown in [27], we can directly expand our SE system to support a class of simple range search [27]. Take a keyword name "Age" with keyword values from 0 to 100 as an example. The path of the leaf node "11-20" is ("0-100", "0-30", "11-20"), and "0-30", "0-10" are simple ranges from level-2 and level-3, respectively.
- Anonymous KP-ABE. Our SE system is built by anonymizing the Rouselakis-Waters KP-ABE scheme [18]. Therefore, our scheme can be easily extended to obtain an unbounded and anonymous KP-ABE scheme in the prime-order group without random oracles, in which an adversary, given a ciphertext, cannot learn any information about the associated attribute set.
- Anonymous hierarchical identity-based encryption (HIBE). The Rouselakis-Waters KP-ABE scheme in [18] can be converted to an HIBE scheme using non-repeating identities, "AND" policies and delegation capabilities [19]. Since our SE scheme can be used to construct an anonymous KP-ABE scheme, it can be further converted to an anonymous HIBE scheme using the same method as in [19].

5 Performance Analysis

We implement our construction of expressive SE in the prime-order group in Charm [39], which is a programming environment for rapid prototyping of cryptographic prim- itives. In this section, we compare the computational cost, communication and storage overhead of our scheme with other existing schemes.

Comparison

Let $|pars|$, $|msk|$, $|CT|$, $|T_{M,\rho}|$, $|M|$ be the sizes of the public parameter, the master private key, the ciphertext, the trapdoor and the access structure, respectively. Let k be the length of the vector corresponding to the ciphertext in [16], l be the number of keywords in an access structure, n be the maximum number of keywords allowed for the system, and m be the size of a keyword set ascribed to a ciphertext. Denote E as an exponentiation operation, P as a pairing operation, χ_1 as the number of elements in $I_{M,\rho} = \{I_1, ..., I_{\chi_1}\}$, χ_2 as $|I_1| + ... + |I_{\chi_1}|$, and χ_3 as the number of primed keywords [14] in a search predicate.

TABLE 2
Comparison of Storage and Communication Overhead

	Public parameter $\|pars\|$	Master private key $\|msk\|$	Trap-door $\|T_{M,\rho}\|$	Cipher-text $\|CT\|$
KSW13 [16]	$2k+3$	$2k+4$	$2k+1+\|M\|$	$2k+1$
LZDLC13 [8]	$n+5$	$n+4$	$2l+\|M\|$	$m+2$
LHZF14 [14]	$n+4$	$n+2$	$3l+\|M\|$	$m+2$
Our Scheme	9	5	$6l+\|M\|$	$5m+2$

We compare our searchable encryption system with the other three known expressive SE schemes [8], [14], [16] in Table 2 which are all constructed over composite order groups. From Table 2, it is not difficult to see that our construction is the only one that supports unbounded num- ber of keywords in the expressive keyword search systems. Note that our scheme is measured in terms of number of elements in prime order groups while the other three schemes are measured in terms of number of elements in composite order groups. According to the analysis in [40][7], in terms of the pairing-friendly elliptic curves, prime order groups have a clear advantage in the parameter sizes over composite order groups.

TABLE 3
Comparison of Computation Overhead.

	KSW13 [16]	LZDLC13 [8]	LHZF14 [14]	Ours
Trap-door	$6k \cdot E$	$4l \cdot E$	$4l \cdot E$	$16l \cdot E + E$
Enc.	$4k \cdot E + E$	$2(m+1) \cdot E$	$(m+2) \cdot E + P$	$7m \cdot E + 2 \cdot E$
Test	$2k \cdot P + P$	$\leq \chi_2 \cdot E + 2\chi_2 \cdot P$	$\leq \chi_2 \cdot E + 2\chi_2 \cdot P + 2\chi_3 \cdot P$	$\leq \chi_2 \cdot E + E + P + 6\chi_2 \cdot P$
Group Order	Composite	Composite	Composite	Prime

In Table 3, we compare the computational costs incurred in the systems from [8], [14], [16] and our system. It is worth noticing that as mentioned in [17], "a Tate pairing on a 1024- bit composite-order elliptic curve is roughly 50 times slower than the same pairing on a comparable prime-order curve,

7. See Table 3 in [40] for the results.

and this performance gap will only get worse at higher security levels". Therefore, although our SE system requires more exponentiation and pairing operations than the other systems, it is far more computationally efficient than the other three schemes.

Experimental Results

We implement our scheme in Charm [39][8], which is a framework developed to facilitate rapid prototyping of cryptographic schemes and protocols. Based on the Python programming language, Charm enables one to implement a cryptographic scheme with very few lines of code, signif- icantly reducing development time. Meanwhile, computa- tionally intensive mathematical operations are implement- ed with native modules, so the overhead due to Python in Charm is less than 1%. Since all Charm routines are designed under the asymmetric groups, our construction is transformed to the asymmetric setting before the imple- mentation. That is, three groups tt, $\hat{tt}$ and tt_1 are used and the pairing $\hat{e}$ is a function from $tt \times \hat{tt}$ to tt_1. Notice that it has been stated in [18] that the assumptions and the security proofs can be converted to the asymmetric setting in a generic way.

We use Charm of version charm-0.43 and Python 3.4 in our implementation. Along with charm-0.43, we install the latest PBC library for underlying cryptographic operations. Our experiments run on an all-in-one desktop computer with Intel Core i7-4785T CPU (4 core 2.20GHz) and 8GB RAM running 64-bit Ubuntu 15.10.

The computational costs of the Setup and sKeyGen algo- rithms are straightforward, and we focus on the computa- tional costs of the Trapdoor, Encrypt and Test algorithms. In our experiments, a set of keywords is generated, of which every keyword contains a generic name such as "Illness", "Position", "Affiliation" and a keyword value such as "Di- abetes", "Doctor", and "City Hospital". For the sake of simple implementation, we use integers to denote keyword values, e.g., a keyword as "Illness = 6" is expressed by "Illness = Diabetes". In this way, we generate a random set of keywords containing 10 to 50 keywords, and use them to encrypt 5,000 documents. We then remove the keyword values in the ciphertexts such that they contain only generic names of keywords like "Illness", "Position", as specified in our concrete construction.

Thereafter, we randomly choose 2 to 10 keywords to form a random access structure. The number of keywords in a searching query is normally less than 10, according to the searching query logs of search engines [41]. The policy tree is formed such that for any interior node the difference on the node number of its left branch and that of its right branch is less than 2. We generate 50 different access policy trees, 10 for each different number of keywords, and create a trapdoor for each policy tree. We also remove the keyword value information from the trapdoors. So the policy tree in a trapdoor contains only keyword names, e.g., (("Illness" AND "Position") OR "Affiliation").

Also, we take each trapdoor to conduct search over the ciphertexts. For any combination of the keyword names in the ciphertext that satisfies the access policy of the trapdoor, our keyword search scheme runs the Test algorithm to further confirm whether it is an exact match.

All these experiments are conducted over 4 different elliptic curves: SS512, MNT159, MNT201 and MNT224, of which SS512 is a supersingular elliptic curve with the bi- linear pairing on it being symmetric Type 1 pairing, and the pairings on the other three curves are asymmetric Type 3 pairings. These four curves provides security levels of 80- bit, 80-bit, 100-bit and 112-bit, respectively. The computation time for the exponentiation and pairing calculation over the four curves are listed in Fig. 2.

ECs(time in ms)	Exp. tt	Exp. $\hat{tt}$	Exp. tt_1	Pairing
SS512	0.194	0.194	0.027	0.881
MNT159	0.068	0.584	0.160	3.148
MNT201	0.101	0.762	0.207	4.194
MNT224	0.131	0.968	0.252	5.169

Fig. 2. Computational costs for the group operations and pairings over different elliptic curves on a desktop with 2.2GHz 4 core CPU.

Fig. 3 shows the computational overhead for generating trapdoors containing 2 keywords to 10 keywords, from which we can see that the computation time for the trapdoor generation is almost linear to the number of keywords as- sociated with the access structure in the trapdoor. The MNT curves with higher security levels have longer computation time, so MNT224 has higher computation cost among all curves. The computation time of SS512 is close to that of MNT224 due to its higher exponentiation cost over tt. The computation time of generating a trapdoor with 10 keywords is only 0.22s for MNT224, which is quite modest for a powerful trapdoor generation centre.

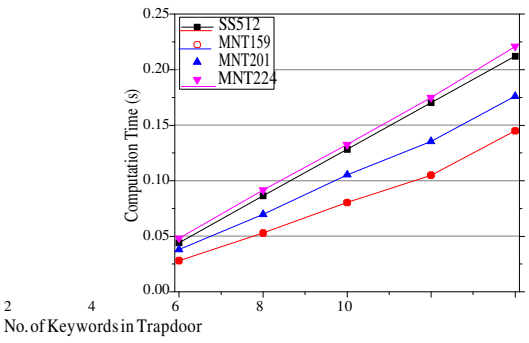


Fig. 3. Computational cost of the Trapdoor algorithm for different curves with respect to number of keywords in trapdoor.

Fig. 4 demonstrates the computation time for the Encrypt algorithm over 10 keywords to 50 keywords. As expected in our analysis, it shows that the computation time is approximately linear to the number of keywords used to generate the ciphertext. The MNT curves with higher secu- rity levels are more expensive in computation cost, while

8. For the explicit information on Charm, please refer to [39]. Note that since it has been clearly shown in [18], [40] that the efficiency of schemes in composite-order groups is much worse than that of schemes in prime-order groups, we will not implement those schemes in composite-order groups (e.g., [8], [14], [16]). In addition, the current version Charm does not support cryptographic schemes in composite- order groups.

the encryption cost of SS512 is much less than that of MNT curves. This is due to the fact that $(4m+1)$ exponentiations are done in $tt\hat{}$ for the total $(7m+2)$ exponentiations (see Table 3). To encrypt a document with 50 keywords using MNT224 curve, the computation time is about 1.6s, which is acceptable for most applications.

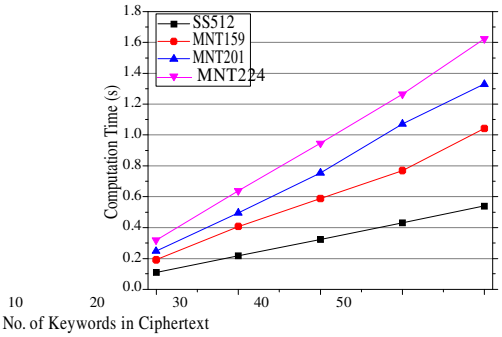


Fig. 4. Computational cost of the Encrypt algorithm for different curves with respect to number of keywords in ciphertext.

The computational cost for the Test algorithm is much more involved. It depends on χ_2, the total number of keywords in all combinations of keywords satisfying the access policy that need to be tried by the cloud server. χ_2 is determined by the access policy and the keywords used to encrypt a document. Fig. 5 shows the relation between the computation time of the Test algorithm and the number of keywords in the access structure of the Trapdoor algorithm. From Fig. 5, it is easy to see that the computation time raises as the number of keywords in the trapdoor and the cipher- text increases. When the trapdoor contains only 2 keywords, the computation time increases quite slowly as keywords in ciphertexts increases. Whilst when the trapdoor has 10 keywords, the computation time grows exponentially as the number of keywords in ciphertexts grows. Among all the curves, SS512 has the best performance, while MNT224 has the highest computational cost. For the 4 curves tested in our experiments, the computation time of searching a document ranges from 50s to 250s for a trapdoor with 10 keywords and a ciphertext with 50 keywords. The computation time can be significantly reduced if keyword search is performed by a powerful cloud server.

6 Conclusions

In order to allow a cloud server to search on encrypted data without learning the underlying plaintexts in the public- key setting, Boneh [7] proposed a cryptographic primitive called public-key encryption with keyword search (PEKS). Since then, considering different requirements in practice, e.g., communication overhead, searching criteria and secu- rity enhancement, various kinds of searchable encryption systems have been put forth. However, there exist only a few public-key searchable encryption systems that support expressive keyword search policies, and they are all built from the inefficient composite-order groups [17]. In this paper, we focused on the design and analysis of public-key searchable encryption systems in the prime-order groups that can be used to search multiple keywords in expressive searching formulas. Based on a large universe key-policy attribute-based encryption scheme given in [18], wepresent- ed an expressive searchable encryption system in the prime- order group which supports expressive access structures expressed in any monotonic Boolean formulas. Also, we proved its security in the standard model, and analyzed its efficiency using computer simulations.

Acknowledgments

This research work is supported by the Singapore Na- tional Research Foundation under the NCR Award Num- ber NRF2014NCR-NCR001-012, and the National Natural Science Foundation of China under the Grant Number 61370027.

References

[1] O. Goldreich and R. Ostrovsky, "Software protection and simula- tion on oblivious rams," *J. ACM*, vol. 43, no. 3, pp. 431–473, 1996.

[2] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in *2000 IEEE Symposium on Security and Privacy, Berkeley, California, USA, May 14-17, 2000*. IEEE Computer Society, 2000, pp. 44–55.

[3] E. Goh, "Secure indexes," *IACR Cryptology ePrint Archive*, vol. 2003, p. 216, 2003.

[4] C. Cachin, S. Micali, and M. Stadler, "Computationally private information retrieval with polylogarithmic communication," in *Advances in Cryptology - EUROCRYPT '99, International Conference on the Theory and Application of Cryptographic Techniques, Prague, Czech Republic, May 2-6, 1999, Proceeding*, ser. Lecture Notes in Computer Science, vol. 1592. Springer, 1999, pp. 402–414.

[5] G. D. Crescenzo, T. Malkin, and R. Ostrovsky, "Single database pri- vate information retrieval implies oblivious transfer," in *Advances in Cryptology - EUROCRYPT 2000, International Conference on the Theory and Application of Cryptographic Techniques, Bruges, Belgium, May 14-18, 2000, Proceeding*, ser. Lecture Notes in Computer Sci- ence, vol. 1807. Springer, 2000, pp. 122–138.

[6] W. Ogata and K. Kurosawa, "Oblivious keyword search," *J. Complexity*, vol. 20, no. 2-3, pp. 356–371, 2004.

[7] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*, ser. Lecture Notes in Computer Science, vol. 3027. Springer, 2004, pp. 506–522.

[8] J. Lai, X. Zhou, R. H. Deng, Y. Li, and K. Chen, "Expressive search on encrypted data," in *8th ACM Symposium on Information, Computer and Communications Security, ASIA CCS '13, Hangzhou, China - May 08 - 10, 2013*. ACM, 2013, pp. 243–252.

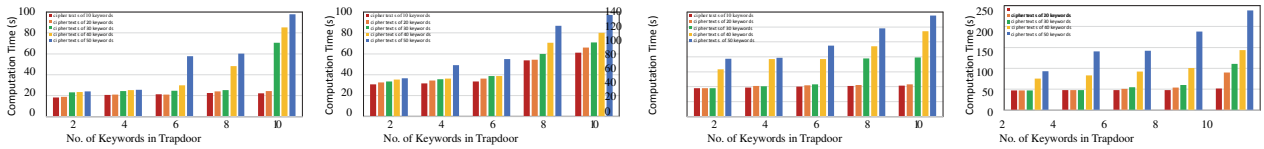
[9] P. Golle, J. Staddon, and B. R. Waters, "Secure conjunctive key- word search over encrypted data," in *Applied Cryptography and Network Security, Second International Conference, ACNS 2004, Yellow Mountain, China, June 8-11, 2004, Proceedings*, ser. Lecture Notes in Computer Science, vol. 3089. Springer, 2004, pp. 31–45.

[10] D. J. Park, K. Kim, and P. J. Lee, "Public key encryption with conjunctive field keyword search," in *Information Security Applications, 5th International Workshop, WISA 2004, Jeju Island, Korea, August 23- 25, 2004, Revised Selected Papers*, ser. Lecture Notes in Computer Science, vol. 3325. Springer, 2004, pp. 73–86.

[11] Y. H. Hwang and P. J. Lee, "Public key encryption with conjunctive keyword search and its extension to a multi-user system," in *Pairing-Based Cryptography - Pairing 2007, First International Con- ference, Tokyo, Japan, July 2-4, 2007, Proceedings*, ser. Lecture Notes in Computer Science, vol. 4575. Springer, 2007, pp. 2–22.

[12] B. Zhang and F. Zhang, "An efficient public key encryption with conjunctive-subset keywords search," *J. Network and Computer Applications*, vol. 34, no. 1, pp. 262–267, 2011.

ISBN 978-93-88808-61-3



(a) SS512 (b) MNT159 (c) MNT201 (d) MNT224

Fig. 5. Experimental results for the Test algorithm over different ellipticcurves.

[13] D. Boneh and B. Waters, "Conjunctive, subset, and range queries on encrypted data," in *Theory of Cryptography, 4th Theory of Cryptography Conference, TCC 2007, Amsterdam, The Netherlands, February 21-24, 2007, Proceedings*, ser. Lecture Notes in Computer Science, vol. 4392. Springer, 2007, pp. 535–554.

[14] Z. Lv, C. Hong, M. Zhang, and D. Feng, "Expressive and secure searchable encryption in the public key setting," in *Information Security - 17th International Conference, ISC 2014, Hong Kong, China, October 12-14, 2014. Proceedings*, ser. Lecture Notes in Computer Science, vol. 8783. Springer, 2014, pp. 364–376.

[15] J. Shi, J. Lai, Y. Li, R. H. Deng, and J. Weng, "Authorized keyword search on encrypted data," in *Computer Security - ESORICS 2014 - 19th European Symposium on Research in Computer Security, Wroclaw, Poland, September 7-11, 2014. Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 8712. Springer, 2014, pp. 419–435.

[16] J. Katz, A. Sahai, and B. Waters, "Predicate encryption supporting disjunctions, polynomial equations, and inner products," *J. Cryp- tology*, vol. 26, no. 2, pp. 191–224, 2013.

[17] D. M. Freeman, "Converting pairing-based cryptosystems from composite-order groups to prime-order groups," in *Advances in Cryptology - EUROCRYPT 2010, 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, French Riviera, May 30 - June 3, 2010. Proceedings*, ser. Lecture Notes in Computer Science, vol. 6110. Springer, 2010, pp. 44–61.

[18] Y. Rouselakis and B. Waters, "Practical constructions and new proof methods for large universe attribute-based encryption," in *2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013*. ACM, 2013, pp. 463–474.

[19] A. B. Lewko and B. Waters, "Unbounded HIBE and attribute- based encryption," in *Advances in Cryptology - EUROCRYPT 2011 - 30th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Tallinn, Estonia, May 15-19, 2011. Proceedings*, ser. Lecture Notes in Computer Science, vol. 6632, 2011, pp. 547–567.

[20] X. Boyen and B. Waters, "Anonymous hierarchical identity-based encryption (without random oracles)," in *Advances in Cryptology - CRYPTO 2006, 26th Annual International Cryptology Conference, Santa Barbara, California, USA, August 20-24, 2006, Proceedings*, ser. Lecture Notes in Computer Science, vol. 4117. Springer, 2006, pp. 290–307.

[21] J. Lai, R. H. Deng, and Y. Li, "Expressive CP-ABE with partially hidden access structures," in *7th ACM Symposium on Information, Compuer and Communications Security, ASIACCS '12, Seoul, Korea, May 2-4, 2012*. ACM, 2012, pp. 18–19.

[22] H. S. Rhee, J. H. Park, W. Susilo, and D. H. Lee, "Improved searchable public key encryption with designated tester," in *Pro- ceedings of the 2009 ACM Symposium on Information, Computer and Communications Security, ASIACCS 2009, Sydney, Australia, March 10-12, 2009*. ACM, 2009, pp. 376–379.

[23] M. Bellare, A. Boldyreva, and A. O'Neill, "Deterministic and efficiently searchable encryption," in *Advances in Cryptology - CRYPTO 2007, 27th Annual International Cryptology Conference, San- ta Barbara, CA, USA, August 19-23, 2007, Proceedings*, ser. Lecture Notes in Computer Science, vol. 4622. Springer, 2007, pp.535–552.

[24] C. Gu, Y. Zhu, and H. Pan, "Efficient public key encryption with keyword search schemes from pairings," in *Information Security and Cryptology, Third SKLOIS Conference, Inscrypt 2007, Xining, China, August 31 - September 5, 2007, Revised Selected Papers*, ser. Lecture Notes in Computer Science, vol. 4990. Springer, 2007, pp. 372–383.

[25] J. Baek, R. Safavi-Naini, and W. Susilo, "Public key encryption with keyword search revisited," in *Computational Science and Its Applications - ICCSA 2008, International Conference, Perugia, Italy, June 30 - July 3, 2008, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 5072. Springer, 2008, pp. 1249–1259.

[26] Q. Tang and L. Chen, "Public-key encryption with registered key- word search," in *Public Key Infrastructures, Services and Applications - 6th European Workshop, EuroPKI 2009, Pisa, Italy, September 10-11, 2009, Revised Selected Papers*, ser. Lecture Notes in Computer Science, vol. 6391. Springer, 2009, pp. 163–178.

[27] M. Li, S. Yu, N. Cao, and W. Lou, "Authorized private keyword search over encrypted data in cloud computing," in *2011 Inter- national Conference on Distributed Computing Systems, ICDCS 2011, Minneapolis, Minnesota, USA, June 20-24, 2011*. IEEE Computer Society, 2011, pp. 383–392.

[28] 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, 2012.

[29] W. Yau, R. C. Phan, S. Heng, and B. 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.

[30] E. Shen, E. Shi, and B. Waters, "Predicate privacy in encryption systems," in *Theory of Cryptography, 6th Theory of Cryptography Conference, TCC 2009, San Francisco, CA, USA, March 15-17, 2009. Proceedings*, ser. Lecture Notes in Computer Science, vol. 5444. Springer, 2009, pp. 457–473.

[31] D. Boneh and M. Franklin, "Identity-based encryption from the weil pairing," in *CRYPTO*, ser. Lecture Notes in Computer Science, vol. 2139. Springer-Verlag, 2001, pp. 213–219.

[32] D. Boneh, X. Boyen, and H. Shacham, "Short group signatures," in *Advances in Cryptology - CRYPTO 2004, 24th Annual International CryptologyConference, Santa Barbara, California, USA, August 15-19, 2004, Proceedings*, ser. Lecture Notes in Computer Science, vol. 3152. Springer, 2004, pp. 41–55.

[33] A. B. Lewko and B. Waters, "Decentralizing attribute-based encryption," in *Advances in Cryptology - EUROCRYPT 2011 - 30th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Tallinn, Estonia, May 15-19, 2011. Proceedings*, ser. Lecture Notes in Computer Science, vol. 6632. Springer, 2011, pp. 568–588.

[34] B. Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," in *Public Key Cryptography - PKC 2011 - 14th International Conference on Practice and Theory in Public Key Cryptography, Taormina, Italy, March 6-9, 2011. Proceedings*, ser. Lecture Notes in Computer Science, vol. 6571. Springer, 2011, pp. 53–70.

[35] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in *Proceedings of the 13th ACM Conference on Computer and Communications Security, CCS 2006, Alexandria, VA, USA, Ioctober 30 - November 3, 2006*. ACM, 2006, pp. 89–98.

[36] R. Ostrovsky, A. Sahai, and B. Waters, "Attribute-based encryption with non-monotonic access structures," in *Proceedings of the 2007 ACM Conference on Computer and Communications Security, CCS 2007, Alexandria, Virginia, USA, October 28-31, 2007*. ACM, 2007, pp. 195–203.

[37] A. B. Lewko, A. Sahai, and B. Waters, "Revocation systems with very small private keys," in *31st IEEE Symposium on Security and Privacy, S&P 2010, 16-19 May 2010, Berleley/Oakland, California, USA*. IEEE Computer Society, 2010, pp. 273–285.

[38] A. Beimel, "Secure schemes for secret sharing and key distri- bution," Ph.D. dissertation, Israel Institute of Technology, Israel Institute of Technology, June 1996.

[39] J. A. Akinyele, C. Garman, I. Miers, M. W. Pagano, M. Rushanan, M. Green, and A. D. Rubin, "Charm: a framework for rapidly prototyping cryptosystems," *J. Cryptographic Engineering*, vol. 3, no. 2, pp. 111–128, 2013.

[40] A. Guillevic, "Comparing the pairing efficiency over composite- order and prime-order elliptic curves," in *Applied Cryptography and Network Security - 11th International Conference, ACNS 2013, Banff, AB, Canada, June 25-28, 2013. Proceedings*, ser. Lecture Notes in Computer Science, vol. 7954. Springer, 2013, pp. 357–372.

[41] L. Yang, Q. Mei, K. Zheng, and D. A. Hanauer, "Query log analysis of an electronic health record search engine," in *Proc. of AMIA Annual Symposium*, 2011, p. 915C924.

APPENDIX A
SYSTEM FRAMEWORK AND SECURITY DEFINITION

Our expressive keyword search system consists of setup algorithm Setup, server key generation algorithm sKeyGen, trapdoor generation algorithm Trapdoor, encryption algo- rithm Encrypt and testing algorithm Test.

- Setup(1^λ) $\rightarrow$ (*pars*, *msk*). Taking the security parameter λ as the input, this setup algorithm outputs the public parameter *pars* and the master private key *msk* for the system. This algorithm is run by the trapdoor centre.
- sKeyGen(*pars*) $\rightarrow$ (sk_s, pk_s). Taking the public parameter *pars* as the input, the server key generation algorithm outputs a public and private key pair for the designated searching server. This algorithm is run by the trapdoor centre.
- Trapdoor(*pars*, pk_s, *msk*, (M, ρ,{ $W_{\rho(i)}$ })) $\rightarrow$ T_M. Taking the public parameter *pars*, the server public key pk_s and an access structure (M, ρ{, $W_{\rho(i)}$}) over the universe of keywords as the input, this trapdoor generation algorithm generates a trapdoor T_M. This algorithm is run by the trapdoor centre.
- Encrypt(*pars*, **W**)$\rightarrow$ CT. Taking the public parameter *pars* and a set of keywords **W** as the input, this encryption algorithm outputs a ciphertext CT. This algorithm is run by the data owner.
- Test(*pars*, sk_s, CT, T_M) $\rightarrow$ 1/0. Taking the public parameter *pars*, the server private key sk_s, a cipher- text CT associated with a keywords set **W** and a trapdoor T_M for an access structure (M, ρ,{$W_{\rho(i)}$ }) as the input, this testing algorithm outputs either 1 when the ciphertext satisfies the access structure of the trapdoor T_M or 0 otherwise. This algorithm is run by the designated server.

We require that a expressive keyword search scheme Π is correct, meaning that for all the sets of keyword- s **W** and access structures M such that M(**W**) = 1, if (*pars*, *msk*) $\leftarrow$ Setup(1^λ), (pk_s, sk_s) $\leftarrow$ sKeyGen(*pars*), $T_M \leftarrow$ Trapdoor(*pars*, pk_s, *msk*, (M, ρ, $W_{\rho(i)}$ }), CT } $\leftarrow$ Encrypt(*pars*, **W**), then Test(*pars*, sk_s, CT, T_M) = 1.

Following the security model introduced in [7], [25], we give the security definition for an expressive keyword search scheme over encrypted data in terms of the semantic security to ensure that such a scheme does not reveal any information about the keyword values in the ciphertext, which we call "indistinguishability against chosen keyword- set attack (IND-CKA)". Formally, we describe the IND-CKA security in the following game between a challenger algorithm B and an adversary algorithm A, where algorithm A is divided into algorithm A_1 (which is assumed to be a designated cloud (searching) server) and algorithm A_2 (which is assumed to be an outside attacker).

1) The security game between algorithm B and algorithm A_1 is to guarantee that the searching server cannot tell which ciphertext encrypts which set of keywords without obtaining the trapdoors for the access structures that can be satisfied by the keywords associated with the ciphertexts. This is because once the server is given a trapdoor that the keyword set in a ciphertext can satisfy, the server will ascertain that this ciphertext contains at least the keywords associated with the access structure in the given trapdoor.
   - Setup. Algorithm B runs the Setup algorithm to obtain the public parameter *pars* and the master private key *msk*. It gives the public parameter *pars* to algorithm A_1 and keeps *msk* to itself. In addition, algorithm B runs the sKeyGen algorithm to obtain a public and private key pair (pk_s, sk_s) for the server. It then gives (pk_s, sk_s) to algorithm A_1.
   - Phase 1. Algorithm A_1 adaptively issues queries to algorithm B for the trapdoors corresponding to the access structures $(M_1, \rho_1, \{W_{\rho_1(i)}\}), \ldots, (M_{q_1}, \rho_{q_1}, \{W_{\rho_{q_1}(i)}\})$. For each $(M_j, \rho_j, \{W_{\rho(i)}\}_j)$ with $j \in [1, q_1]$, algorithm B runs the Trapdoor algorithm, and sends T_{M_j} to algorithm A_1.
   - Challenge. Algorithm A_1 outputs two sets of keyword $\mathbf{W}^*_0, \mathbf{W}^*_1$ of the same size with the restriction that $\mathbf{W}_0$ and $\mathbf{W}_1$ satisfy none of the queried trapdoors. Algorithm B selects a random bit $\beta \in \{0, 1\}$, runs the Encrypt algorithm on $\mathbf{W}^*_\beta$ to obtain the challenge ciphertext CT^*, and then forwards CT^* to algorithm A_1.
   - Phase 2. Algorithm A_1 continues issuing queries to algorithm B for the trapdoors corresponding to the access structures $(M_{q_1+1}, \rho_{q_1+1}, \{W_{\rho_{q_1+1}(i)}\}), \ldots, (M_q, \rho_q, \{W_{\rho_q(i)}\})$ with the restriction that any $(M_j, \rho_j, \{W_{\rho_j(i)}\})$ for $j \in [q+1, q]$ can be satisfied by neither $\mathbf{W}_0$ nor $\mathbf{W}_1$.
   - Guess. Algorithm A_1 outputs its guess $\beta' \in \{0, 1\}$ and wins the game if $\beta' = \beta$.
2) The security game between algorithm B and algorithm A_2 is to ensure that the outsider attacker who has not obtained the searching server's private key cannot determine the set of keyword values associated with the ciphertext even though the attacker gets the trapdoors over the access structures satisfied by the keywords associated with the ciphertexts. This is because the server's public key is embedded in the trapdoors such that no one can determine whether a trapdoor matches the keyword set of a ciphertext without the server's private key.

ISBN 978-93-88808-61-3

- Setup. Algorithm B runs the Setup algorithm to obtain the public parameter *pars* and the master private key *msk*. It gives the public parameter *pars* to algorithm A_2 and keeps *msk* to itself. Also, algorithm B runs the sKey- Gen algorithm to obtain a public and private key pair (pk_s, sk_s) for the server. It then gives pk_s to algorithm A_2 and keeps sk_s to itself.
- Phase 1. Algorithm A_2 adaptively issues queries to algorithm B for the trapdoors cor- responding to the access structures $(M_1, \rho_1, \{W_{\rho_1(i)}\}), \ldots, (M_{q_1}, \rho_{q_1}, \{W_{\rho_{q_1}(i)}\})$. For each M_j with $j \in [1, q_1]$, algorithm B runs the trapdoor generation algorithm Trapdoor, and sends T_{M_j} to algorithm A_2.
- Challenge. Algorithm A_2 outputs two sets of keyword $\mathbf{W}_0^*, \mathbf{W}_1^*$ of the same size. Algorithm B selects a random bit $\beta \in \{0, 1\}$, runs the Encrypt algorithm on $\mathbf{W}_\beta^*$ to obtain the challenge ciphertext CT^*, and then gives CT^* to algorithm A_2.
- Phase 2. Algorithm A_2 continues issuing queries to algorithm B for the trapdoors corresponding to the access structures $(M_{q_1+1}, \rho_{q_1+1}, \{W_{\rho_{q_1+1}(i)}\}), \ldots, (M_q, \rho_q, \{W_{\rho_q(i)}\})$.
- Guess. Algorithm A_2 outputs its guess $\beta' \in \{0, 1\}$ and wins the game if $\beta' = \beta$.

For $A \in \{A_1, A_2\}$, an expressive keyword search system Π is IND-CKA secure if the advantage function referring to the security game $\text{Game}^{\text{IND}}_{\Pi, A}$

$$\mathbf{Adv}^{\text{IND}}_{\Pi,A}(\lambda) \stackrel{\text{def}}{=} \Pr[\beta = \beta'] \quad \cdot$$

is negligible in the security parameter λ for any probabilistic polynomial-time (PPT) adversary algorithm A.

In addition, an expressive keyword search system is said to be selectively IND-CKA secure[9] if an Init stage is added before the Setup phase where algorithm A commits to the challenge keyword sets $\mathbf{W}_0^*, \mathbf{W}_1^*$ which it aims to attack.

APPENDIX B
SECURITY PROOF OF THEOREM 1

Proof. The proof is divided into two parts, depending on the role of the adversary. In the first part, the adversary is assumed to be an outside attacker, and in the second part, the adversary is assumed to be the server.

In terms of the first part of the proof, we prove it via a sequence of games, where game Game_0 is the same as the original game, and game Game_1 is the same as Game_0 except that the trapdoors might be generated in a different way. We finish the proof by showing that if there exists an outside adversary algorithm A that can distinguish game Game_1 from game Game_0, then we can build a challenger algorithm B that solves the decisional BDH assumption.

9. Note that selective IND-CKA security is weaker than IND-CKA security, but it is a useful tool in security reduction and is widely used in the cryptographic systems.

- Algorithm A gives algorithm B two challenge keyword sets $\mathbf{W}_0^* = \{W_{0,1}^*, \ldots, W_{0,m}^*\}$ and $\mathbf{W}_1^* = \{W_{1,1}^*, \ldots, W_{1,m}^*\}$.
- Setup. Algorithm B runs the Setup algorithm to generate the public parameter and the master private key as required, and sets the public and private key pair for the server as (g^a, a). Also, algorithm B selects a random bit $\beta \in \{0, 1\}$.
- Phase 1. Since algorithm B knows the master private key, it easily outputs the trapdoor on any access structures as required. If algorithm A issues a trap- door generation query on an access structure that can be satisfied by $\mathbf{W}_\beta^*$, algorithm B computes $T = g^c$, $T' = g^b$, $T_{i,2} = H(\hat{e}(g, g)^{abc}) g^{d_1 d_2 t_{i,1} + d_3 d_4 t_{i,2}}$, and generates the other elements of the trapdoor as in the Trapdoor algorithm.
- Challenge. Algorithm B runs the Encrypt algorithm on $\mathbf{W}_\beta^*$ to obtain the challenge ciphertext CT^*, and gives CT^* to algorithm A.
- Phase 2. The same as that in Phase 1.
- Guess. Algorithm A output a guess β' for β.

On the one hand, if $Z = \hat{e}(g, g)^{abc}$, then algorithm A's view of this simulation is identical to the original game. On the other hand, if Z is randomly chosen from tt_1, then algorithm A's advantage is nil. Therefore, if algorithm A can discern game Game_1 from game Game_0 with a non-negligible probability, algorithm B has a non-negligible probability in breaking the decisional BDH problem.

Concerning the second part of the proof, we prove the security using a sequence of games. For simplicity, we remove the access structure from the ciphertext, and denote $CT^* = (C^*, D^*, \{D^*, E^*_{i,1}, E^*_{i,2}, F^*_{i,1}, F^*_{i,2})\}_{i \in [1,m]})$ as the challenge ciphertext given to the adversary during an attack in the real world. Let Z be a random element of tt_1, and $\{Z_{i,1}\}$, $\{Z_{i,2}\}$ be random elements of tt. We define the following games which differ on the type of the challenge ciphertext is given by the challenger to the adversary.

- Game_0: The challenge ciphertext is $CT_0 = (C^*, D^*, \{D^*, E^*_{i,1}, E^*_{i,2}, F^*_{i,1}, F^*_{i,2})\}_{i \in [1,m]})$.
- Game_1: The challenge ciphertext is $CT^* = (C^* \cdot Z, D^*, \{D^*, E^*_{i,1}, E^*_{i,2}, F^*_{i,1}, F^*_{i,2})\}_{i \in [1,m]})$.
- Game_2: The challenge ciphertext is $CT_2 = (C^* \cdot Z, D^*, (D^*, Z_{1,1}, E^*_{1,1}, E^*_{1,2}, F^*_{1,1}, F^*_{1,2}), \{(D^*, E^*_{i,1}, E^*_{i,2}, F^*_{i,1}, F^*_{i,2})\}_{i \in [2,m]})$.
- ……
- Game_{m+1}: The challenge ciphertext is $CT^*_m = (C^* \cdot Z, D^*, \{(D^*, Z^*_{i,1}, E^*_{i,2}, F^*_{i,1}, F^*_{i,2})\}_{i \in [1,m]})$.
- Game_{m+2}: The challenge ciphertext is $CT_{m+2} = (C^* \cdot Z, D^*, (D^*, Z_{1,1}, E^*_{1,1}, E^*_{1,2}, F^*_{1,1}, F^*_{1,2}), \{(D^*, Z^*_{i,1}, E^*_{i,2}, F^*_{i,1}, F^*_{i,2})\}_{i \in [2,m]})$.
- ……
- Game_{2m+1}: The challenge ciphertext is $CT_{2m+1} = (C^* \cdot Z, D^*, \{(D^*_i, Z_{i,1}, E^*_{i,2}, Z^*_{i,1}, F^*_{i,2})\}_{i \in [1,m]})$.

To complete the proof, we will show that the games Game_0, Game_1, ..., Game_{2m+1} are computationally indis- tinguishable from each other.

Lemma 1. Under the $(q-2)$ assumption, the advantage for a polynomial time adversary that can distinguish between the games Game_0 and Game_1 is negligible.

ISBN 978-93-88808-61-3

Proof. Assume that there is an adversary algorithm A that can distinguish Game_0 from Game_1. Then we can build a challenger algorithm B that can solve the $(q-2)$ problem.

- Init. Algorithm A gives algorithm B two challenge keyword sets $\mathbf{W}^* = \{W_{0,1}^*, ..., W_{0,m}^*\}$ and $\mathbf{W} = \{W_{1,1}^*, ..., W_{1,m}^*\}$.
- Setup. In order to generate the public system parameter, algorithm B implicitly sets $\alpha = xy$. Then algorithm B randomly chooses $\beta \in \{0, 1\}$, d_1, d_2, d_3, d_4, $\tilde{u}$, $\tilde{h} \in Z_p$, and computes the public parameter $pars = (g, u, h, w, g_1, g_2, g_3, g_4, \hat{e}(g^x, g^y))$ as follows.

$$g = g,\ w = g^{x},\ g_1 = g^{d_1},\ g_2 = g^{d_2},$$
$$g_3 = g^{d_3},\ g_4 = g^{d_4},\ u = g^{\tilde{u}} \cdot \prod_{i\in[m]} g^{y/b_i^2},$$
$$h = g^{\tilde{h}} \cdot \prod_{i\in[m]} (g^{xz/b_i}) \cdot \prod_{i\in[m]} (g^{y/b_i^2})^{-W_{\beta,i}^*}.$$

- Phase 1 and Phase 2. Algorithm B has to create the trapdoors for the access structures $(\mathbf{M}, \rho, \{\rho(i)\})$ required by algorithm A that are not satisfied by either $\mathbf{W}^0$ or $\mathbf{W}^1$. Since $\mathbf{W}$ does not satisfy $(\mathbf{M}, \rho)$, there exists a vector $\vec{w} = (w_1, ..., w_n) \in Z_p^n$ such that $w_1 = 1$ and $\mathbf{M}_i \cdot \vec{w} = 0$ for all $i \in [l]$ such that $\rho(i) \in \mathbf{W}^*_{\beta}$. Algorithm B computes w using linear algebra. The vector y shared is implicitly set as $\vec{y} = xy\vec{w} + (0, \tilde{y}_2, ..., \tilde{y}_n)^{\perp}$, where $\tilde{y}_2, ..., \tilde{y}_n \in Z_p$. This is a properly distributed vector with first component as $xy = \alpha$ and the other components being uniformly random in Z_p. As a result, for each row $i \in [l]$, the share is

$$v_i = \mathbf{M}_i \cdot \vec{y} = xy(\mathbf{M}_i \cdot \vec{w}) + (\mathbf{M}_i \cdot (0, \tilde{y}_2, ..., \tilde{y}_n)^{\perp})$$
$$= xy(\mathbf{M}_i \cdot \vec{w}) + \tilde{v}_i.$$

As mentioned above, for each row i for which $\rho(i) \in \mathbf{W}^*_{\beta}$, $\mathbf{M}_i \cdot \vec{w} = 0$. In this case, $v_i = \tilde{v}_i = \mathbf{M}_i \cdot (0, \tilde{y}_2, ..., \tilde{y}_n)^{\perp}$, which is known to algorithm B, so algorithm B randomly chooses $t_i \in Z_p$, and outputs $\{T_{i,1}, T_{i,2}, T_{i,3}, T_{i,4}, T_{i,5}, T_{i,6}\}$ as in the Trapdoor algorithm.

For each row $i \notin \mathbf{W}^*_{\beta}$, algorithm B randomly chooses $\tilde{t}_{i,1}, \tilde{t}_{i,2} \in Z_p$, and implicitly sets

$$T_{i,1} = g^{v_i} w^{d_1 d_2 \tilde{t}_{i,1} + d_3 d_4 \tilde{t}_{i,2}}$$
$$\cdot \prod_{j\in[n]} (g^{x z_j})^{\frac{\mathbf{M}_i \cdot \vec{w}}{\rho(i) - W_{\beta,j}} \cdot \tilde{t}_{i,1} d_1 d_2} \cdot \prod_{j\in[n]} (g^{x z_j})^{\frac{\mathbf{M}_i \cdot \vec{w}}{\rho(i) - W_{\beta,j}^*} \cdot \tilde{t}_{i,2} d_3 d_4}.$$

$$T_{i,2} = g^{d_1 d_2 \tilde{t}_{i,1} + d_3 d_4 \tilde{t}_{i,2}} = (g^y)^{-\mathbf{M}_i \cdot \vec{w}}$$
$$\cdot \prod_{j\in[m]} (g^{xzb_j})^{\frac{\mathbf{M}_i \cdot \vec{w}}{\rho(i) - W_{\beta,j}} \cdot \tilde{t}_{i,1} d_1 d_2},$$
$$\cdot \prod_{j\in[m]} (g^{xzb_j})^{\frac{\mathbf{M}_i \cdot \vec{w}}{\rho(i) - W_{\beta,j}^*} \cdot \tilde{t}_{i,2} d_3 d_4}$$
$$(g^{\quad}) \qquad \cdot g^{\quad}.$$

$$T_{i,3} = ((u^{\rho(i)} h)^{-t_{i,1}})^{-d_2} = (g^y)^{(\mathbf{M}_i \cdot \vec{w})(\rho(i) u + \tilde{h}) d_2}$$
$$\cdot (g^{xzb_j})^{\frac{\tilde{\quad}}{\rho(i) - \beta,j} \vec{d}_2}$$
$$\cdot \prod_{j\in[m]} (g^{(xz)^2 \tilde{b}_k / b_j})^{\frac{-(\mathbf{M}_i \cdot \vec{w})}{\rho(i) - W_{\beta k}} d_2}$$
$$\cdot \prod_{(j,k)\in[m,m]} (g^{y^2/b_j^2})^{(\mathbf{M}_i \cdot w)(\rho(i) - W_{\beta,j})^* d_2}$$
$$\cdot \prod_{j\in[m]} \prod_{jf=k} (g^{\frac{xzb_k}{b_j}})^{\frac{-(\mathbf{M}_i \cdot \vec{w})(\rho(i) - W_{\beta,j}^*)}{\rho(i) - W_{\beta k}}} d_2$$
$$\cdot ((u^{\rho(i)} h)^{-\tilde{t}_{i,1}})^{-d_2}.$$

Since $T_{i,3}$, $T_{i,4}$, $T_{i,5}$, $T_{i,6}$ have the term $(u^{\rho(i)} h)^{-t_{i,1}}$ in common, and d_1, d_2, d_3 and d_4 are known to algo- rithm B, algorithm B can simply compute $T_{i,4}$, $T_{i,5}$, $T_{i,6}$ as $T_{i,3}$. Thus, algorithm B successfully responds to algorithm A's trapdoor queries.

- Challenge. To generate a challenge ciphertext, algorithm B implicitly sets $\mu = z$ from the $q-2$ assumption, and $z_i = b_i$ for every $i \in [m]$. Notice that these parameters are properly distributed since $z, b_1, ..., b_q$ are information theoretically hidden from the view of algorithm A. In addition, algorithm B

ISBN 978-93-88808-61-3

$\tilde{t}_{i,1}$ $2d_3d_4$ $\tilde{y}$

On the one hand, if $Z = \hat{e}(g, g)^{xyz}$, then algorithm A's view of this simulation is identical to the original game. On the other hand, if Z is randomly chosen from tt_1, then algorithm A's advantage is nil. Therefore, if algorithm A can distinguish game Game_1 from game Game_0 with a non-negligible probability, algorithm B has a non-negligible probability in breaking the $(q-2)$ assumption.

$\tilde{w}^{-1}$ z_m

Lemma 2. Under the decisional linear assumption, the advantage for a polynomial time adversary that can distinguish between the games Game_{m+1} and Game_m for $m \in [1, m]$ is negligible.

Proof. Assuming that there is an adversary algorithm A that can distinguish Game_m from Game_{m+1}, we can build a challenger algorithm B to solve the decisional linear problem.

- Init. Algorithm A gives algorithm B two challenge keyword sets $\mathbf{W}_0 = \{W^*_{0,1}, ..., W^*_{0,m}\}$, $\mathbf{W}_1 = \{W^*_{1,1}, ..., W^*_{1,m}\}$.
- Setup. In order to generate the public system parameter, algorithm B implicitly sets $d_1 = x_2$, $d_2 = x_1$. Then algorithm B randomly chooses $d_3, d_4, \beta \in \{0, 1\}$, α, y, $\tilde{w} \in Z_p$, and computes the public parameter $pars = (g, u, h, w, g_1, g_2, g_3, g_4, \hat{e}(g, g)^{\alpha})$.

$m+m+1$

$$g = g, \quad w = g^{\tilde{w}}, \quad g_1 = g^{x_2}, \quad g_2 = g^{x_1},$$
$$g_3 = g^{x_3}, \; g_4 = g^{x_4}, \; u = g^{x_2\alpha},$$
$$h = g^{-x_2\alpha W^*_{\beta,m}} g^{y}, \quad \hat{e}(g, g)^{\alpha} = \hat{e}(g, g)^{\alpha}.$$

- Phase 1 and Phase 2. In order to create a trapdoor for an access structure (M, ρ) required by algorithm A that is satisfied by neither $\mathbf{W}^*_0$ nor $\mathbf{W}^*_1$, algorithm B performs as follows. It randomly chooses $\vec{y} = (\alpha, y_2, ..., y_n)^{\perp}$ where $y_2, ..., y_n \in Z_p$. Also, it randomly chooses $t_{1,1}, t_{1,2}, ..., t_{l,1}, t_{l,2} \in Z_p$. For each $i \in [l]$, algorithm B sets $v_i = M_i \cdot \vec{y}$,

$$\tilde{t}_{i,1} = \frac{t_{i,1}\alpha(\rho(i) - W^*_{\beta,m})}{x_2\alpha(\rho(i) - W^*_{\beta,m}) + y'} + \frac{y x_1 t_{i,1}}{d_3 d_4(\rho(i) - W^*_{\beta,m})x_2 + y}.$$

$\tilde{t} = t$ $_{i,2}$ $_{i,2}$

Then it outputs the trapdoor as

$$T_{i,1} = g^{v_i}(g^{x_1 t_{i,1}} g^{t_{i,2} d_3 d_4})^{\tilde{w}} = g^{v_i} w^{d_1 d_2 \tilde{t}_{i,1} + d_3 d_4 \tilde{t}_{i,2}},$$
$$T_{i,2} = g^{x_1 t_{i,1}} g^{t_{i,2} d_3 d_4} = g^{x_1 x_2 \tilde{t}_{i,1} + \tilde{t}_{i,2} d_3 d_4}$$
$$= g^{d_1 d_2 \tilde{t}_{i,1} + d_3 d_4 \tilde{t}_{i,2}},$$
$$T_{i,3} = (g^{x_1})^{-\alpha(\rho(i) - W^*_{\beta,m})t_{i,1}} = ((u^{\rho(i)}h)^{\tilde{t}_{i,1}})^{-d_2},$$
$$= (g^{x_2})^{-\alpha(\rho(i) - W^*_{\beta,m})t_{i,1}} = ((u^{\rho(i)}h)^{\tilde{t}_{i,1}})^{-d_1},$$
$$T_{i,5} = (g^{x_1})^{-\alpha(\rho(i) - W^*_{\beta,m})t_{i,2}} = ((u^{\rho(i)}h)^{\tilde{t}_{i,2}})^{-d_2},$$
$$T_{i,6} = (g^{x_2})^{-\alpha(\rho(i) - W^*_{\beta,m})t_{i,2}} = ((u^{\rho(i)}h)^{\tilde{t}_{i,2}})^{-d_1}.$$

- Challenge. To generate a challenge ciphertext, algorithm B implicitly sets $s_{m,1} = x_3$, $z_m = x_3 + x_4$ from the decisional linear assumption. In addition, algorithm B randomly chooses μ, $s_{1,1}, ..., s_{m-1,1}, s_{1,2}, ..., s_{m,2} \in Z_p$, $z_1, ..., z_{m-1} \in Z_p$. Thus, algorithm B can calculate the challenge ciphertext as follows.

1) For $i = m$, algorithm B outputs

$$C^* = \hat{e}(g, g)^{\alpha\mu}, \quad D^* = g^{\mu},$$

$$D_m = w^{-\mu y}(u^{\beta,m}h)^{-\mu y} = w^{-\mu y}Z,$$
$$E^*_{m,1} = g_1^{z_m - s_1} = g^{x_2 x_4},$$
$$E^*_{m,2} = g_2^{s_{m,1}} = g^{x_1 x_3},$$
$$F^*_{m,1} = Z^{d_3} \cdot g_3^{-s_{m,2}}, \; F^*_{m,2} = g_4^{s_{m,2}}.$$

2) For any $i \in [m-1]$, algorithm B outputs

$$D^* = w^{-\mu}(u^{W^*_{\beta,i}}h)^{z_i}, \quad E^*_{i,1} = g_1^{z_i - s_{i,1}},$$
$$E^*_{i,2} = g_2^{s_{i,1}}, \; F^*_{i,1} = g_3^{z_i - s_{i,2}}, \; F^*_{i,2} = g_4^{s_{i,2}}.$$

- Guess. Algorithm A output a guess β' for β.

On the one hand, if $Z = g^{x_3+x_4}$, then algorithm A's view of this simulation is identical to the original game. On the other hand, if Z is randomly chosen from tt, then algorithm A's advantage is nil. Therefore, if algorithm A can distinguish game Game_m from game Game_{m+1} with a non-negligible probability, algorithm B has a non-negligible probability in breaking the decisional linear assumption.

Lemma 3. Under the decisional linear assumption, the advantage for a polynomial time adversary that can distinguish between the games Game_t and Game_t for $m' \in [1, m]$ is negligible.

proof. This proof follows almost the same as that of Lemma 2, except that the simulation is done over the parameters g_3 and g_4 instead of g_1 and g_2.

This completes the proof of Theorem 1.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

1

A Comprehensive Study on Social Network Mental Disorders Detection via Online Social Media Mining

Naveen

M.Tech Scholar,
CSE Department
Malla Reddy College of Engineering
naveenit@gmail.com

Abstract—The explosive growth in popularity of social networking leads to the problematic usage. An increasing number of social network mental disorders (SNMDs), such as Cyber-Relationship Addiction, Information Overload, and Net Compulsion, have been recently noted. Symptoms of these mental disorders are usually observed passively today, resulting in delayed clinical intervention. In this paper, we argue that mining online social behavior provides an opportunity to actively identify SNMDs at an early stage. It is challenging to detect SNMDs because the mental status cannot be directly observed from online social activity logs. Our approach, new and innovative to the practice of SNMD detection, does not rely on self-revealing of those mental factors via questionnaires in Psychology. Instead, we propose a machine learning framework, namely, *Social Network Mental Disorder Detection (SNMDD)*, that exploits features extracted from social network data to accurately identify potential cases of SNMDs. We also exploit multi-source learning in SNMDD and propose a new SNMD-based Tensor Model (STM) to improve the accuracy. To increase the scalability of STM, we further improve the efficiency with performance guarantee. Our framework is evaluated via a user study with 3126 online soc ial network users. We conduct a feature analysis, and also apply SNMDD on large-scale datasets and analyze the characteristics of the three SNMD types. The results manifest that SNMDD is promising for identifying online social network users with potential SNMDs.

Index Terms—Tensor factorization acceleration, online social network, mental disorder detection, feature extraction.

1 INTRODUCTION

With the explosive growth in popularity of social network- ing and messaging apps, online social networks (OSNs) have become a part of many people's daily lives. Most research on social network mining focuses on discovering the knowledge behind the data for improving people's life. While OSNs seemingly expand their users' capability in increasing social contacts, they may actually decrease the face-to-face interpersonal interactions in the real world. Due to the epidemic scale of these phenomena, new terms such as Phubbing (Phone Snubbing) and Nomophobia (No Mobile Phone Phobia) have been created to describe those who cannot stop using mobile social networking apps.

- *Hong-Han Shuai is with the Department of Electrical Computer Engineering, National Chiao Tung University, No. 1001, University Road, Hsinchu, Taiwan. E-mail: hhshuai@nctu.edu.tw*
- *Chih-Ya Shen is with the Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan. E-mail: chihya@cs.nthu.edu.tw*
- *De-Nian Yang and Ming-Syan Chen are with the Research Cen- ter of Information Technology Innovation, Academia Sinica, No. 128, Sec. 2, Academia Road, Taipei, 11529 Taiwan. Email: {dnyang, mschen}@citi.sinica.edu.tw.*
- *Yi-Feng Lan is with the Graduate Institute of Educational Psychology and Counseling, Tamkang University, Taiwan. E-mail: carolyflan@gmail.com.*
- *Wang-Chien Lee is with the Department of Computer Science and Engineering, Pennsylvania State University, PA, USA. E-mail: wlee@cse.psu.edu.*
- *Philip S. Yu is with the Department of Computer Science, University of Illinois at Chicago, IL, USA. E-mail: psyu@uic.edu.*
- *Ming-Syan Chen is also with the Department of Electrical Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan.*

In fact, some social network mental disorders (SNMDs), such as Information Overload and Net Compulsion [1], have been recently noted.[1] For example, studies point outthat 1 in 8 Americans suffer from problematic Internet use[2]. Moreover, leading journals in mental health, such as the American Journal of Psychiatry [2], have reported that the SNMDs may incur excessive use, depression, social with- drawal, and a range of other negative repercussions.

Indeed, these symptoms are important components of diagnostic criteria for SNMDs [3] e.g., excessive use of social networking apps – usually associated with a loss of the sense of time or a neglect of basic drives, and with- drawal – including feelings of anger, tension, and/or de- pression when the computer/apps are inaccessible. SNMDs are social-oriented and tend to happen to users who usually interact with others via online social media. Those with SNMDs usually lack offline interactions, and as a result seek cyber-relationships to compensate. Today, identification of potential mental disorders often falls on the shoulders of supervisors (such as teachers or parents) passively. How- ever, since there are very few notable physical risk factors, the patients usually do not actively seek medical or psycho- logical services. Therefore, patients would only seek clinical interventions when their conditions become very severe.

However, a recent study shows a strong correlation be- tween suicidal attempt and SNMDs [4], which indicates that adolescents suffering from social network addictions have a much higher risk of suicidal inclination than non-addictive

1. http://phys.org/news/2015-09-social-media-impacts-mental- well-being.html
2. http://netaddiction.com/faqs/

ISBN 978-93-88808-61-3

users. The research also reveals that social network addic- tion may negatively impact emotional status, causing higher hostility, depressive mood, and compulsive behavior. Even more alarming is that the delay of early intervention may seriously damage individuals' social functioning. In short, it is desirable to have the ability to actively detect potential SNMD users on OSNs at an early stage.

Although previous work in Psychology has identified several crucial mental factors related to SNMDs, they are mostly examined as standard diagnostic criteria in survey questionnaires. To automatically detect potential SNMD cases of OSN users, extracting these factors to assess users' online mental states is very challenging. For example, the extent of loneliness and the effect of disinhibition of OSN users are not easily observable.[3] Therefore, there is a need to develop new approaches for detecting SNMD cases of OSN users. We argue that mining the social network data of individuals as a complementary alternative to the con- ventional psychological approaches provides an excellent opportunity to *actively identify* those cases at an early stage. In this paper, we develop a machine learning framework for detecting SNMDs, which we call *Social Network Mental Disorder Detection (SNMDD)*.

Specifically, we formulate the task as a semi-supervised classification problem to detect three types of SNMDs [1]:
i) Cyber-Relationship Addiction, which shows addictive be- havior for building online relationships; ii) Net Compulsion, which shows compulsive behavior for online social gaming or gambling; and iii) Information Overload, which is related to uncontrollable surfing. By exploiting machine learning techniques with the ground truth obtained via the current diagnostic practice in Psychology [1], we extract and analyze the following crucial categories of features from OSNs: 1) social comparison, 2) social structure, 3) social diversity,
4) parasocial relationships, 5) online and offline interaction ratio, 6) social capital, 7) disinhibition, 8) self-disclosure, and 9) bursting temporal behavior. These features capture important factors or serve as proxies for SNMD detection. For example, studies manifest that users exposed to positive posts from others on Facebook with similar background are inclined to feel malicious envy and depressed due to the social comparison [36]. The depression leads users to disorder behaviors, such as information overload or net compulsion. Therefore, we first identify positive newsfeeds and then calculate the profile similarity and relation famil- iarity between friends. As another example, a parasocial relationship is an asymmetric interpersonal relationship, i.e., one party cares more about the other, but the other does not. This asymmetric relationship is related to loneliness, one of the primary mental factors pushing users with SNMDs to excessively access online social media [5]. Therefore, we extract the ratio of the number of actions to and from friends of a user as a feature. In this paper, the extracted features are carefully examined through a user study.

Furthermore, users may behave differently on different OSNs, resulting in inaccurate SNMD detection. When the data from different OSNs of a user are available, the accu- racy of the SNMDD is expected to improve by effectively integrating information from multiple sources for model training. A na¨ıve solution that concatenates the features from different networks may suffer from the curse of dimen- sionality. Accordingly, we propose an *SNMD-based Tensor Model (STM)* to deal with this multi-source learning problem in SNMDD. Advantages of our approach are: i) the novel *STM* incorporates the SNMD characteristics into the tensor model according to Tucker decomposition; and ii) the tensor factorization captures the structure, latent factors, and cor- relation of features to derive a full portrait of user behavior. We further exploit CANDECOMP/PARAFAC (CP) decom- position based *STM* and design a stochastic gradient descent algorithm, i.e., *STM-CP-SGD*, to address the efficiency and solution uniqueness issues in traditional Tucker decom- position. The convergence rate is significantly improved by the proposed second-order stochastic gradient descent algorithm, namely, *STM-CP-2SGD*. To further reduce the computation time, we design an approximation scheme of the second-order derivative, i.e., Hessian matrix, and provide a theoretical analysis.

The contributions of this paper are summarized below.

- Today online SNMDs are usually treated at a late stage. To actively identify potential SNMD cases, we propose an innovative approach, new to the current practice of SNMD detection, by mining data logs of OSN users as an early detection system.
- We develop a machine learning framework to detect SNMDs, called *Social Network Mental Disorder Detec- tion (SNMDD)*. We also design and analyze many im- portant features for identifying SNMDs from OSNs, such as disinhibition, parasociality, self-disclosure, etc. The proposed framework can be deployed to provide an early alert for potential patients.
- We study the *multi-source learning* problem for SNMD detection. We significantly improve the efficiency and achieve the solution uniqueness by CP decom- position, and we provide theoretical results on non- divergence. By incorporating SNMD characteristics into the tensor model, we propose *STM* to better extract the latent factors from different sources to improve the accuracy.
- We conduct a user study with 3126 users to evaluate the effectiveness of the proposed SNMDD frame- work. To the best of our knowledge, this is the first dataset crawled online for SNMD detection. Also, we apply SNMDD on large-scale real datasets, and the results reveal interesting insights on network structures in SNMD types, which can be of interest to social scientists and psychologists.

The rest of this paper is organized as follows. Section 2 surveys the related work. Section 3 presents *SNMDD*, fo- cusing on feature extraction. Section 4 presents the proposed *STM* for multi-source learning and the acceleration of tensor decomposition with the theoretical results. Section 5 reports a user study, various analyses, and the experimental results. Finally, Section 6 concludes this paper.

3. The online disinhibition effect is a loosening (or complete aban- donment) of social restrictions and inhibitions that would otherwise be present in normal face-to-face interaction during interactions with others on the Internet.

ISBN 978-93-88808-61-3

2 Related Work

Internet Addiction Disorder (IAD) is a type of behavior addiction with the patients addicted to the Internet, just like those addicting to drugs or alcohol [3]. Many research works in Psychology and Psychiatry have studied the important factors, possible consequences, and correlations of IAD [10], [40], [41], [42]. King et al. [40] investigate the problem of sim- ulated gambling via digital and social media to analyze the correlation of different factors, e.g., grade, ethnicity. Baumer et al. [10] report the Internet user behavior to investigate the reason of addiction. Li et al. [41] examine the risk factors related to Internet addiction. Kim et al. [42] investigate the association of sleep quality and suicide attempt of Internet addicts. On the other hand, recent research in Psychology and Sociology reports a number of mental factors related to social network mental disorders. Research indicates that young people with narcissistic tendencies and shyness are particularly vulnerable to addiction with OSNs [6], [7]. However, the above research explores various negative im- pacts and discusses potential reasons for Internet addiction. By contrast, this paper proposes to automatically identify SNMD patients at the early stage according to their OSN data with a novel tensor model that efficiently integrate heterogeneous data from different OSNs.

Research on mental disorders in online social networks receives increasing attention recently [43], [44], [45]. Among them, content-based textual features are extracted from user-generated information (such as blog, social media) for senti- ment analysis and topic detection. Chang et. al [43] employ an NLP-based approach to collect and extract linguistic and content-based features from online social media to iden- tify Borderline Personality Disorder and Bipolar Disorder patients. Saha et al. [44] extract the topical and linguistic features from online social media for depression patients to analyze their patterns. Choudhury et al. [45] analyze emo- tion and linguistic styles of social media data for Major De- pressive Disorder (MDD). However, most previous research focuses on individual behaviors and their generated textual contents but do not carefully examine the structure of social networks and potential Psychological features. Moreover, the developed schemes are not designed to handle the sparse data from multiple OSNs. In contrast, we propose a new multi-source machine learning approach, i.e., STM, to extract proxy features in Psychology for different diseases that require careful examination of the OSN topologies, such as Cyber-Relationship Addiction and Net Compulsion.

Our framework is built upon support vector machine, which has been widely used to analyze OSNs in many areas [11], [12]. In addition, we present a new tensor model that not only incorporates the domain knowledge but also well estimates the missing data and avoids noise to properly handle multi-source data. Caballero et al. [8] estimate the probability of mortality in ICU by modeling the probability of mortality as a latent state evolving over time. Zhao et al. [9] propose a hierarchical learning method for event de- tection and forecasting by first extracting the features from different data sources and then learning via geographical multi-level model. However, the SNMD data from different OSNs may be incomplete due to the heterogeneity. For example, the profiles of users may be empty due to the privacy issue, different functions on different OSNs (e.g., game, check-in, event), etc. We propose a novel tensor-based approach to address the issues of using heterogeneous data and incorporate domain knowledge in SNMD detection.

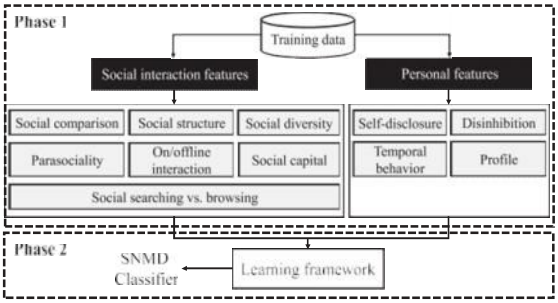


Fig. 1. The SNMDD framework.

3 Social Network Mental Disorder Detection

In this paper, we aim to explore data mining techniques to detect three types of SNMDs [1]: 1) *Cyber-Relationship (CR) Addiction*, which includes the addiction to social net- working, checking and messaging to the point where social relationships to virtual and online friends become more important than real-life ones with friends and families; 2) *Net Compulsion (NC)*, which includes compulsive online social gaming or gambling, often resulting in financial and job-related problems; and 3) *Information Overload (IO)*, which includes addictive surfing of user status and news feeds, leading to lower work productivity and fewer social inter- actions with families and friends offline.

Accordingly, we formulate the detection of SNMD cases as a classification problem. We detect each type of SNMDs with a binary SVM. In this study, we propose a two-phase framework, called *Social Network Mental Disorder Detection (SNMDD)*, as shown in Figure 1. The first phase extracts various discriminative features of users, while the second phase presents a new SNMD-based tensor model to derive latent factors for training and use of classifiers built upon Transductive SVM (TSVM) [13]. Two key challenges exist in design of SNMDD: i) we are not able to directly extract mental factors like what have been done via questionnaires in Psychology and thus need new features for learning the classification models;[4] ii) we aim to exploit user data logs from multiple OSNs and thus need new techniques for integrating multi-source data based on SNMD characteris- tics. We address these two challenges in Sections 3.1 and 4, respectively.

Feature Extraction

We first focus on extracting discriminative and informative features for design of SNMDD. This task is nontrivial for the following three reasons.

1. **Lack of mental features.** Psychological studies have shown that many mental factors are related to SNMDs, e.g., low self-esteem [3], loneliness [14]. Thus, questionnaires are designed to reveal those factors for SNMD detection. Some

4. Additional issues in feature extraction will be detailed later.

parts of Psychology questionnaire for SNMDs are based on the subjective comparison of mental states in online and offline status, which cannot be observed from OSN logs. For example:

Q1. How often do you feel depressed, moody, or nervous when you are off-line, which goes away once you are back online?
Q2. How often do you prefer the excitement of the Internet to intimacy with your partner?

Consider Q1. The feel of depression and nervousness offline can not be observed online. To tackle this problem, we have to leverage the knowledge from Psychology, such as withdrawal or relapse patterns, and exploit some proxy features extracted from online social activity logs to approxi- mate them. For Q2, the preference of excitement of the Inter- net to intimacy with users' partners is an important question for SNMD detection. As it is difficult to directly observe these factors from data collected from OSNs, psychiatrists are not able to directly assess the mental states of OSN users under the context of online SNMD detection.

2. Heavy users vs. addictive users. To detect SNMDs, an intuitive idea is to simply extract the usage (time) of a user as a feature for training SNMDD. However, this feature is not sufficient because i) the status of a user may be shown as "online" if she does not log out or close the social network applications on mobile phones, and ii) heavy users and addictive users all stay online for a long period, but heavy users do not show symptoms of anxiety or depression when they are not using social apps. How to distinguish them by extracting discriminative features is critical.

3. Multi-source learning with the SNMD characteristics. As we intend to exploit user data from different OSNs in SNMDD, how to extract complementary features to draw a full portrait of users while considering the SNMD characteristics into the tensor model is a challenging problem.

To address the first two challenges, we identify a number of effective features as proxies to capture the mental states of users, e.g., self-esteem [3] and loneliness [14].[5] The goal is to distinguish users with SNMDs from normal users. Two types of features are extracted to capture the social inter- action behavior and personal profile of a user. Due to the space constraint, some of the above features are presented in Appendix A. It is worth noting that each individual feature cannot precisely classify all cases, as research shows that exceptions may occur. Therefore, it is necessary to exploit multiple features to effectively remove exceptions.

Social Interaction Features

We first extract a number of *social interaction features* to capture the user behavior on social media.

Social comparison based features (SComp) Although most literature indicates that the majority of the newsfeed up- dates is positive, recent studies manifest that users who are exposed to positive posts from others on Facebook are in- clined to feel envy and depressed due to social comparison [38]. The social comparison leads to SNMDs according to Festinger's theory, which states that many people usually have a strong motivation to evaluate their own opinions and abilities by implicitly or explicitly comparing with others in similar backgrounds, especially when the reference in comparison to the physical world is not specific. The situa- tion becomes increasingly serious because status exchanges among friends are now very convenient via various online social networks.

Envy usually appears after comparisons, and two kinds of envy, i.e., benign envy and malicious envy, exist in Psychology [36]. The experience of benign envy leads to a moving-up motivation aiming at improving one's own position, whereas the experience of malicious envy pro- duces a pulling-down motivation and depression. Malicious envy is incurred from the comparison among close friends with similar backgrounds and states, and it usually leads to SNMDs, such as information overload or net compul- sion, because a person in this case usually feels pressure and tends to frequently check the updated status of the corresponding friends. A teenager student in this case may seek online games or gambles as alternatives for acquiring the sense of accomplishment. By contrast, benign envy is usually generated from distant friends with different back- grounds and rarely leads to SNMDs.

Therefore, for malicious envy, we first exploit the ex- isting techniques of emotional signal processing [17] to identify positive newsfeeds and then calculate the profile similarity and relation familiarity between friends. Specifi- cally, let $N_p(i, j)$ and $s(i, j)$ denote the number of positive newsfeeds that user j receives from i and the similarity on backgrounds between user i and j, respectively. For user j, the weighted number of positive newsfeeds based on similarity can be derived as

$$\frac{\sum_{i \in N(j)} [s(i, j) N_p(i,j)]}{\sum_{i \in N(j)} s(i, j)}, \quad (1)$$

where $N(j)$ is the set of neighbors of user j. Moreover, the weighted numbers of positive newsfeeds based on familiar- ity can also be derived in a similar manner by substituting the similarity function with the familiarity function as an additional proxy feature for the social comparison.

Social structure based features (SS) In Sociology, each person in a social network belongs to one of the following three types of social roles: influential users, structural holes, and normal users. An influential user is the one with a huge degree and many mentions and shares (retweets) [28]. On the other hand, weaker connecting paths between groups are structure holes in OSNs, and researchers have demon- strated that structural holes usually have timely access to important information, e.g., trade trend, job opportunities, which usually leads to social success. Therefore, the users with their roles as structural holes are more inclined to suf- fer from information overload for newsfeeds because they enjoy finding and sharing new and interesting information to various friends.

According to the above observations, we exploit the state-of-the-art approach [27] to quantify users' tendencies of being structural holes. Specifically, given n users and m communities, let $F \in R^{n \times m}$ denote the community indicator matrix, where $f_{ij} = 1$ if a user i is assigned to the j-th community, and 0 otherwise. Let $\mathbf{f}^i$ denote the i-th row vector of F. By embedding the harmonic function to learn

5. The third challenge is addressed in Section 4.

the community indicator matrix, the difference between the value of f^i and the averaged value of its neighbors $\frac{1}{d_i}\sum_{(v_i,v_j)\in E} f^j$ is required to be minimized, because neighbors are usually within similar communities. Therefore, the following minimization problem is formulated to detect structural hole spanners.

$$\min_F \|F - D^{-1}AF\|_{2,1} \quad (2)$$

$$\text{s.t. } F^T F = I_m, \quad (3)$$

where A and D are respectively the adjacency and degree matrices, and $\|X\|_{2,1}$ is the $l_{2,1}$ norm of X, which is the sum of the Euclidean norms of the columns of the matrix X. The structural hole spanners correspond to the ones with small $\|f^i\|$. Compared with social capital based features, the structural hole feature considers the community structure (global), while social capital features only examine the ego networks (local). On the other hand, we also extract the network topology based features, i.e., closeness centrality, betweenness centrality, eigenvector centrality, information centrality, flow betweenness, the rush index, as social structure based features for detecting SNMDs. For example, flow betweenness indicates how much information has been propagated through the node, which relates to information overload. Moreover, eigenvector centrality is a measure of the influence of a node in a network, and the score is similar to the pagerank, i.e., connections to high-scoring neighbors are inclined to increase the score of a node. Therefore, the scores of unpopular users are usually small and correlated to Cyber-Relationship Addiction.

Social diversity based features (SDiv) Researchers have observed that diversity improves the depth of people thinking for both majority or minority [35]. For example, a person with a more diverse background and many friends is less inclined to suffer from SNMDs because she is often supported by friends and thereby rarely feels lonely and isolated (two important factors correlated to SNMDs) [34]. Therefore, the impact of social network diversity is increasingly important and inspires us to incorporate them for effective SNMD detection. Specifically, the diversities of nationality, racial, ethical, religious, and education can be extracted as social diversity based features with Shannon index H as the diversity index, i.e.,

$$H = -\sum_{i=1}^{N_t} p_i \ln p_i, \quad (4)$$

where p_i and N_t are the proportion of users' friends belonging to the i-th type of attributes and the total number of types, respectively. The value H increases when the number of types N_t grows. Moreover, Shannon diversity index also increases when there is a more significant *evenness*. In other words, the diversity index is maximized when all type of attributes are of the equal quantities.

Parasocial relationship (PR). Research shows that the mental factor of loneliness is one of the primary reasons whythe users with SNMDs excessively access online social media [5]. As the loneliness of an OSN user is hard to measure, we exploit the parasocial relationship, an asymmetric interpersonal relationship between two people where one party cares more about the other but the other does not, to capture loneliness (as studies show that they are correlated [15]). The feature of parasocial relationship is represented as $|a_{out}|/|a_{in}|$, where $|a_{out}|$ and $|a_{in}|$ denote the number of actions a user takes to friends and the number of actions friends take to the user, respectively.[6] As the ratio increases, the extent of parasocial relationship also grows.

Due to the space constraint, other social interaction features are presented in Appendix A.1.

Personal Features

Temporal behavior features (TEMP). *Relapse* is the state that a person is inclined to quickly revert back to the excessive usage of social media after an abstinence period, while *tolerance* is the state that the time spent by a person with SNMDs tends to increase due to the mood modification effect.[7] It is worth noting that the above two mental states have been exploited to evaluate clinical addictions [14]. We aim to use them to distinguish *heavy users* and *addictive users* because heavy users do not suffer from relapse and tolerance in use of OSNs. An issue arising here is how to assess relapse and tolerance quantitatively.

It is observed that the use of social media by an SNMD patient is usually in the form of *intermittent bursts* [3]. Therefore, given a stream of a user's activities on an OSN, e.g., "likes", "comments", "posts", we exploit Kleinberg's burst detection algorithm [16], which is based on an infinite Markov model, to detect periods of the user's activities as bursty and non-bursty periods. The bursty period refers to a period during which the activities significantly increase. A bursty period is modeled as a bursty state q_1 in the Markov model, while a non-bursty period is correspondingly modeled as a normal state q_0. The burst detection algorithm finds a state transition sequence q for each user to divide the corresponding log (stream of activities) into bursty and non-bursty periods. Specifically, let $x = (x_1, x_2, \ldots, x_n)$ denote a sequence of n time intervals between $n+1$ consecutive activities, with theintervals distributed according to a density function, such as $f_{i_t}(x_t) = \alpha_{i_t} e^{-\alpha_{i_t} x_t}$, where α_{i_t} is either α_0 or α_1, and α_0 and α_1 are parameters that correspond to the normal and burst states, respectively, $\alpha_1 > \alpha_0$. A time interval x_t is in a burst state q_1 if $f_0(x_t) < f_1(x_t)$. Otherwise, it is in a normal state q_0. However, simply deciding the state sequence q based on this criteria results in numerous small periods. Therefore, a cost $\tau(q_i, q_j)$ is associated with a state transition from q_i to q_j to filter out noises and to ensure that each bursty or non-bursty period is sufficiently long. Therefore, the remaining issue is to find an optimal state-transition sequence q to minimize the following cost function [16],

$$c(q|x) = \sum_{t=1}^{n-1} \tau(q_i, q_{i+1}) + \sum_{t=1}^{n} (-\ln f_{i_t}(x_t)),$$

where $\tau(q_i, q_{i+1}) = 0$ if the state q_i and q_{i+1} is the same. $\tau(q_i, q_{i+1})$ is $\gamma \ln n$ otherwise, where γ is an algorithm parameter larger than 0. Notice that the state sequence that minimizes the cost depends on 1) how easy it is to jump from one state to another and 2) how well it is to comply to the rates of arrivals. After identifying the bursts, we

6. The actions include like, comment, and post in our work.

7. A patient may need to spend more time on social media to reach the happiness/excitement than before.

measure their intensity (the number of activities within a burst) and length (the time period of a burst) as the proxy features for *relapse* and *tolerance*, respectively. The(average, median, standard deviation, maximum, minimum) of both the burst intensity and burst length are included in our feature set, because they capture the characteristic of bursts. For instance, the standard deviation of the burst length for SNMD patients is usually larger than that for heavy users since heavy users constantly use OSNs, whereas the users with SNMDs increase the usage time due to tolerance.

Due to the space constraint, other personal features are presented in Appendix A.2.

4 Multi-Source Learning with Tensor Decomposition Acceleration

Many users are inclined to use different OSNs, and it is expected that data logs of these OSNs could provide enriched and complementary information about the user behavior. Thus, we aim to explore multiple data sources (i.e., OSNs) in SNMDD, in order to derive a more complete portrait of users' behavior and effectively deal with the data sparsity problem. To exploit multi-source learning in SNMDD, one simple way is to directly concatenate the features of each person derived from different OSNs as a huge vector. However, the above approach tends to miss the correlation of a feature in different OSNs and introduce interference. Thus, we explore tensor techniques which have been used increasingly to model multiple data sources be- cause a tensor can naturally represent multi-source data. We aim to employ tensor decomposition to extract common latent factors from different sources and objects. Based on tensor decomposition on $\mathcal{T}$, we present a *SNMD-based Tensor Model (STM)* in previous work [47], which enables U to incorporate important characteristics of SNMDs, such as the correlation of the same SNMD sharing among close friends.[8] Finally, equipped with the new tensor model, we conduct semi-supervised learning to classify each user by exploiting Transductive Support Vector Machines (TSVM) in Appendix B. In the following, the problem definition, notation expla- nation, and brief introduction are first presented for better reading.

Problem Definition and Notation Explanation

Given D SNMD features of N users extracted from M OSN sources, we construct a three-mode tensor $\mathcal{T} \in \mathbb{R}^{N \times D \times M}$, where each element $t_{ijk} \in \mathcal{T}$ represents the j-th feature of user i in source k. The objective here is to extract the latent features for each user with tensor composition from $\mathcal{T}$. Here scalars are denoted by lowercase letters, e.g., u, while vectors are denoted by boldface lowercase letters, e.g., $\mathbf{u}$. Matrices are represented by boldface capital letters, e.g., $\mathbf{U}$, and tensors are denoted by calligraphic letters, e.g., $\mathcal{T}$. The i-th row and the j-th column of a two-dimensional matrix $\mathbf{U}$ are respectively denoted by $\mathbf{u}_{i:}$ and $\mathbf{u}_{:j}$.

Tucker decomposition and CANDECOMP/PARAFAC (CP) decomposition have been widely used for extracting the latent features. In the following, we first briefly intro- duce Tucker decomposition.

8. Note that D does not capture the social correlations among friends.

Tucker Decomposition

Tucker decomposition [46] of a tensor $\mathcal{T} \in \mathbb{R}^{N \times D \times M}$ is defined as follows.

$$\mathcal{T} = \mathcal{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}, \tag{5}$$

where $\mathbf{U} \in \mathbb{R}^{N \times R}$, $\mathbf{V} \in \mathbb{R}^{D \times S}$ and $\mathbf{W} \in \mathbb{R}^{M \times T}$ are latent matrices. R, S, and T are parameters to be set according to different criteria [46]. The 1-mode product of $\mathcal{C} \in \mathbb{R}^{R \times S \times T}$ and $\mathbf{U} \in \mathbb{R}^{N \times R}$, denoted by $\mathcal{C} \times_1 \mathbf{U}$, is a matrix with size NST, where each element $(\mathcal{C} \times_1 \mathbf{U})_{nst} = \sum_{r=1}^{R} c_{rst} u_{rn}$. Given the input tensor matrix $\mathcal{T}$ that consists of the features of all users from every OSN, Tucker decomposition derives $\mathcal{C}$, U, V, and W to meet the above equality on $\mathcal{T}_{ndm}$ for every n, d, and m, where $\mathcal{C}$ needs to be diagonal, and U, V, and W are required to be orthogonal [46]. Matrix U effectively estimates a deficit feature (e.g., a missing feature value unavailable due to privacy setting) of an OSN from the corresponding feature of other OSNs, together with the features of other users with the similar behavior.

CP Decomposition

Although Tucker decomposition is flexible and general, it is difficult to interpret the latent features intuitively from the decomposed matrices due to complicated interactions among them. Also, ensuring identifiability is fundamental and important for tensor decomposition. Moreover, the model parameters are encouraged to be uniquely recov- ered given the observed statistics, i.e., the decomposition yields a unique solution. For Tucker decomposition, the identifiability needs to satisfy complicated criteria, e.g., the structured sparsity and symmetry constraints on the core tensor, and sparsity constraints on the inverse factors of the tensor decomposition [48]. In contrast, the latent features obtained by CANDECOMP/PARAFAC (CP) decomposition [33] are much easier to interpret due to the rank-1 compo- nent factorization of CP and its intrinsic axis property from parallel proportional profiles. Moreover, Kruskal criterion on the rank of tensors provides a sufficient condition of the identifiability. Most importantly, its computational com- plexity is much lower than Tucker decomposition, thereby allowing us to analyze SNMDs for large-scale OSNs.

Specifically, CANDECOMP/PARAFAC (CP) decomposi- tion of a tensor $\mathcal{T} \in \mathbb{R}^{N \times D \times M}$ is defined as follows.

$$\sum_{r=1}^{R} \mathbf{U}_{:r} \circ \mathbf{V}_{:r} \circ \mathbf{W}_{:r} \approx \mathcal{T}, \tag{6}$$

where $\circ$ denotes the vector outer product, and R is a positive integer representing the dimensionality of U, V, and W, i.e., $\mathbf{U}_{:r} \in \mathbb{R}^{N}$, $\mathbf{V}_{:r} \in \mathbb{R}^{D}$, and $\mathbf{W}_{:r} \in \mathbb{R}^{m}$, for $r = 1, \ldots, R$. The space of variables in CP decomposition is comprised of the elements of U, V, and W. The inner product of third-order tensors $\mathcal{X}$ and $\mathcal{Y}$ is defined as $\sum_i \sum_j \sum_k X_{ijk} Y_{ijk}$. The objective function of CP decomposition is to find U, V, and W such that the decomposition is close to $\mathcal{T}$ (i.e., the difference is minimized). Each element $\mathbf{U}_{:r} \circ \mathbf{V}_{:r} \circ \mathbf{W}_{:r}$ is a rank-one tensor, and $\mathbf{U}_{i:}$ represents the SNMD feature tensor of user i. Compared to Tucker decomposition, the core tensor $\mathcal{C}$ in CP decomposition has been simplified, and thus the number of parameters required to be estimated

ISBN 978-93-88808-61-3

in Equation (6) is much smaller. Moreover, the solution is unique in CP decomposition but not unique in Tucker decomposition. Equipped with CP decomposition, the objective function $L(T, U, V, W)$ is

$$\frac{1}{2}\|T - \sum_{r=1}^{R} U_{:r} \circ V_{:r} \circ W_{:r}\|^2 + \frac{\lambda_1}{2} tr(U^T L_a U) + \frac{\lambda_2}{2}\|U\|^2, \quad (7)$$

where $tr(\cdot)$ denotes the matrix trace, the Frobenius norm of a tensor T is defined as $\|T\| = \sqrt{< T, T >}$, and λ_1 and λ_2 are parameters controlling the contribution of each part during the above collaborative factorization. The Laplacian matrix L_a of the weighted adjacency matrix A is defined as $D - A$, where D is a diagonal matrix with the entries $d_{ii} = \sum_j a_{ij}$. L first minimizes the decomposition error, i.e., $\|T - \sum_{r=1}^{R} U_{:r} \circ V_{:r} \circ W_{:r}\|^2$ for T. Moreover, the term that minimizes $\|U\|^2$ is to derive a more concise latent feature matrix and avoid overfitting. The proposed *STM* is different from the conventional tensor models in the second term of Eq. (7), where important characteristics of SNMDs are incorporated. For example, the probability of finding CR cases around a CR patient is higher than that around a non- CR user due to the loneliness propagation [15]. That is, CR users usually feel lonely and are more likely to establish friendships in cyberspace with other users with similar behavior. Since the nearby nodes with a great quantity of interactions tend to be the same (either CR or non-CR), it is envisaged that the distance between $u_{i:}$ and $u_{j:}$ will be small if the edge weight of the edge connecting user i and user j, i.e., $a_{i,j}$ in the adjacency matrix A, is sufficiently large. Therefore, a regularization (smoothing) term $\frac{\lambda_1}{2} tr(U^T L_a U)$ is included in the model to achieve the above goal. Due to the space constraint, the details of deriving $\frac{\lambda_1}{2} tr(U^T L_a U)$ are presented in Appendix C.

Stochastic Gradient-Descent Algorithm

Notice that CP decomposition is non-convex. Fortraditional gradient descent algorithms [25], the learning step size η and the initial values on U, V, and W are very sensi- tive and need to be carefully determined. Otherwise, the algorithm is inclined to diverge, consequently failing to find the decomposition solution. To address this issue, we design a new stochastic gradient-descent algorithm with low computational complexity to guarantee the solution convergence.

We present a stochastic gradient-descent algorithm for CP decomposition of the SNMD-based Tensor Model, namely, *SGD-CP-STM*, to iteratively improve each element in the matrices according to the corresponding gradient. Specifically, let $\Gamma(\cdot, V, W)$ be a matrix obtained from by contracting V and W, i.e.,

$$\Gamma(\cdot, V, W)_{ir} = \sum_j \sum_k T_{ijk} V_{jr} W_{kr}, \quad (8)$$

where $\Gamma(\cdot, V, W) \in R^{N \times R}$ (the same as U). The following lemma first derives the gradient of each iteration.

Lemma 1. *The gradient of* L *with regard to* U, *i.e.,* $\nabla_U L(T, U, V, W)$, *is equal to*

$$-\Gamma(\cdot, V, W) + U(\Gamma(V, W) + \lambda_2 I_R) + \lambda_1 L_a U,$$

where $\Gamma(V, W)$ *is the Hadamard product of* $V^T V$ *and* $W^T W$, *i.e.,* $\Gamma(V, W)_{ij} = (V^T V)_{ij} (W^T W)_{ij}$, *and* I_R *is the identity matrix of size* R.

Proof. The objective function $L(T, U, V, W)$ is comprised of three terms, and the derivative of $\frac{\lambda_2}{2}\|U\|^2$ with regard to U is $\lambda_2 I_R$. For the first term, the CP gradient can be solved by the following equation according to [26].

$$\nabla_U \frac{1}{2}\|T - \sum_{r=1}^{R} U_{:r} \circ V_{:r} \circ W_{:r}\|^2 = -\Gamma(\cdot, V, W) + U\Gamma(V, W). \quad (9)$$

For the second term, i.e., $\frac{\lambda_1}{2} tr(U^T L_a U)$, the gradient for U is

$$\nabla_U \frac{\lambda_1}{2} tr(U^T L_a U) = \frac{\lambda_1}{2}(L_a + L_a^T)U. \quad (10)$$

If the weighted adjacency matrix A is symmetric, Equation (10) can be further simplified to $\lambda_1 L_a U$, and $\nabla_U L(T, U, V, W)$ is equal to

$$-\Gamma(\cdot, V, W) + U(\Gamma(V, W) + \lambda_2 I_R) + \lambda_1 L_a U. \quad (11)$$

The theorem follows. □

Therefore, the stochastic gradient descent algorithm up- dates U at the t-th iteration as follows.

$$U^{(t)} = U^{(t-1)} - \eta^{(t)}(-\Gamma^{(t-1)}(\cdot, V^{(t-1)}, W^{(t-1)}) + U^{(t-1)}(\Gamma(V^{(t-1)}, W^{(t-1)}) + \lambda_2 I_R) + \lambda_1 L_a U^{(t-1)}). \quad (12)$$

Based on Eq. (9), the gradient for V and W can be derived in the similar way as follows:

$$\nabla_V L(T, U, V, W) = -\Gamma(U, \cdot, W) + V\Gamma(U, W)$$
$$\nabla_W L(T, U, V, W) = -\Gamma(U, V, \cdot) + W\Gamma(U, V).$$

Note that $V^{(t)}$ and $W^{(t)}$ are also updated similarly in each iteration.

Acceleration of Convergence Rate

It has been widely recognized that the performance is poor when a constant learning step size η is adopted. If the learn- ing step size is too large, it is inclined to skip the optimal solution. In contrast, if the learning step size is too small, the convergence rate to the optimal solution becomes unac- ceptably slow. To avoid the above issue, we further propose a second-order stochastic gradient descent algorithm for CP decomposition, namely, *STM-CP-2SGD*. The second-order stochastic gradient descent (2SGD) considers the second- order information by adaptively assigning the learning rate as the inverse of the Hessian matrix in objective function L to guide the searching direction. Since the inversion of the full Hessian matrix

$$\begin{bmatrix} \nabla_U \nabla_U L & \nabla_U \nabla_V L & \nabla_U \nabla_W L \\ \nabla_W \nabla_U L & \nabla_W \nabla_V L & \nabla_W \nabla_W L \end{bmatrix}$$

is computationally expensive, its block-diagonal parts are alternatively used as an approximation of the inverse of the Hessian matrix [29]. Therefore, we update $U^{(t)}$ as follows.

$$U^{(t)} = U^{(t-1)} - \eta^{(t)} (\nabla^2_{U^{(t-1)}} L)^{-1} \nabla_{U^{(t-1)}} L. \quad (13)$$

ISBN 978-93-88808-61-3

To calculate $\nabla^2_U L$, we take a derivative of Eq. (11) with respect to U and obtain

$$\nabla^2_U L(T, U, V, W) = \frac{d(U(\Gamma(V, W)+\lambda_2 I_R))}{dU} + \lambda_1 \frac{d(L_a U)}{dU} \quad (14)$$

We find the derivatives by exploiting the relationship between the Kronecker product and the vec operator (vec- torizing matrices by stacking its columns) as follows. For the first term in Eq. (14), we have

$$\frac{d(U(\Gamma(V, W)+\lambda_2 I_R))}{dU} = \frac{d(\mathrm{vec}(U(\Gamma(V, W)+\lambda_2 I_R))}{d\mathrm{vec}(U)}. \quad (15)$$

Since $\mathrm{vec}(U(\Gamma(V, W) + \lambda_2 I_R)) = \mathrm{vec}(I_N U(\Gamma(V, W) + \lambda_2 I_R)) = (\Gamma(V, W) + \lambda_2 I_R)^T \otimes I_N \mathrm{vec}(U)$, where $\otimes$ is Kronecker product, we have

$$\frac{d(U(\Gamma(V, W)+\lambda_2 I_R))}{dU} = (\Gamma(V, W)+\lambda_2 I_R)^T \otimes I_N. \quad (16)$$

We find the second term in Eq. (14) in a similar manner.

$$\frac{d(L_a U)}{dU} = \frac{d\mathrm{vec}(L_a U)}{d\mathrm{vec}(U)}. \quad (17)$$

Since $\mathrm{vec}(L_a U) = \mathrm{vec}(L_a U I_R) = I_R \otimes L_a \mathrm{vec}(U)$, we have

$$\frac{d(L_a U)}{dU} = I_R \otimes L_a. \quad (18)$$

Substituting Eqs. (16) and (18) into Eq. (14), $\nabla^2_U L(T, U, V, W)$ is equal to

$$(\Gamma(V, W)+\lambda_2 I_R)^T \otimes I_N + \lambda_2 I_R \otimes L_a = (\Gamma(V, W)+\lambda_2 I_R)^T \oplus \lambda_2 L_a, \quad (19)$$

where $\oplus$ is the Kronecker sum. Therefore, weupdate U at the t-th iteration as follows, i.e., $U^{(t)}$ is equal to

$$U^{(t-1)} - \eta^{(t)}((\Gamma(V^{(t-1)}, W^{(t-1)})+\lambda_2 I_R)^T \oplus \lambda_2 L_a)^{-1} \nabla_{U^{(t-1)}} L. \quad (20)$$

Compared with Eq. (12), here the update of U includes a new term, i.e., $((\Gamma(V, W)+\lambda_2 I_R)^T \oplus \lambda_2 L_a)^{-1}$, representing the adaptive learning step size. Previous studies show that the number of iterations for 2SGD to reach the optimum is much smaller than that of SGD [49]. More specifically, given a second-order convex function, 2SGD requires only one iteration because it derives the optimal step size to the opti- mum point. To properly choose the $\eta^{(t)}$, several approaches can be applied,e.g., damped Newton method [50]. Here, we employ cross validation on the training dataset to find the initial value $\eta^{(0)}$ with Adagrad accordingly. Moreover, we update $V^{(t)}$ and $W^{(t)}$ as follows.

$$V^{(t)} = V^{(t-1)} - \eta^{(t)}(\Gamma(U^{(t-1)}, W^{(t-1)})^T \otimes I)^{-1} \nabla_{V^{(t-1)}} L,$$
$$W^{(t)} = W^{(t-1)} - \eta^{(t)}(\Gamma(U^{(t-1)}, V^{(t-1)})^T \otimes I)^{-1} \nabla_{W^{(t-1)}} L.$$

2SGD, which exploits a Hessian matrix, outperforms SGD because the optimal step size is equal to the inverse of the Hessian matrix when the surface is approximated as a quadratic plane. *STM-CP-2SGD* further utilizes the block diagonal components to approximate the Hessian matrix for acceleration. Notice that if the Hessian matrix is not invertible, a nonnegative diagonal matrix with negligible- valued elements can be added to the original matrix to produce a positive definite matrix [30].

Theoretical Results

In the following, we first derive the computational com - plexity of the above two algorithms. Afterward, we prove that *STM-CP-2SGD* always converges to the solution of CP decomposition. More specifically, let $|T|$ denote the number of nonzero elements in T .

Lemma 2. *The computational complexity of* STM-CP-SGD *is* $O((N + D + M)R^2 + N^2R + |T^{(t)}|R)$.

Proof. For V and W, the complexity of *STM-CP-SGD* is $O((D + M)R^2 + |T^{(t)}|R)$ for each update, where $O((D + M)R^2)$ is for computing $V\Gamma(U, W)$ and $W\Gamma(U, V)$, and $O(|T^{(t)}|R)$ is to find $\Gamma^{(t)}(U, , W)$ and $\Gamma^{(t)}(U, V,)$. For U, the complexity of *STM-CP-SGD* is $O(NR^2 + |T^{(t)}|R + N^2R)$ for each update since the time to find $L_a U$ is $O(N^2R)$. Therefore, the computational complexity of *STM-CP-SGD* is $O((N+D+M)R^2+N^2R+|T^{(t)}|R)$. The lemma follows. □

Lemma 3. *The computational complexity of* STM-CP-2SGD *is* $O((N + D + M)R^2 + N^2R + |T^{(t)}|R)$.

Proof. Compared with *STM-CP-SGD*, *STM-CP-2SGD* needs to additionally derive the inverse of the Hessian matrix. To update V, since $(\Gamma(U, W)^T \otimes I)^{-1} = (\Gamma(U, W))^{-1} \otimes I^{-1}$, we only have to additionally calculate the inverse of $\Gamma(U, W)$, and thus the computational complexity is $O(R^3)$, as well as for the update of W. On the other hand, for the update of U, efficiently calculating the inverse of the Kronecker sum, i.e., $\Gamma(V, W)^T \oplus L_a \in \mathbb{R}^{RN \times RN}$, has been studied in solving Sylvester equation in control system, and it can be approximated in $O(N^2 + R^2)$ [31]. Therefore, the computational complexity is still $O((N+D+M)R^2+N^2R+|T^{(t)}|R)$. The lemma follows. □

The worst-case computational complexity of *STM-CP-2SGD* is the same as *STM-CP-SGD*. Nevertheless, the com- putation of *STM-CP-2SGD* is slightly more complicated than that of *STM-CP-SGD* because *STM-CP-2SGD* needs to compute the inverse of the Hessian matrix. Neverthelss, as shown in the experimental results, the convergence rate of *STM-CP-2SGD* is much faster than that of *STM-CP-SGD*. Given the same start point, *STM-CP-2SGD* can adaptively control the learning step size to approach the optimal so- lution and thus requires much fewer iterations than that of *STM-CP-SGD* with a constant learning step size.

Assume that the Frobenius norm of T is bounded by a constant, i.e., $\|T\| \leq C$, we have the following theorem.

Theorem 1. *Given any initial solution or step size,* STM-CP-2SGD *does not diverge.*

Proof. STM-CP-2SGD updates U, V, W by finding the first and second derivatives. Consider the t-th update of $U^{(t)}$.
We first fix the values of $V^{(t-1)}$ and $W^{(t-1)}$ and set Eq. (11) as zero to find U^*, i.e.,

$$U(\Gamma(V^{(t-1)}, W^{(t-1)})+\lambda_2 I_R)+\lambda_1 L_a U = T_{(:,:,)}(V^{(t-1)}, W^{(t-1)}). \quad (21)$$

Since Eq. (21) is a Sylvester-type equation, the optimal solution of U when fixing $V^{(t-1)}$ and $W^{(t-1)}$, denoted as U^*, can be derived as

$$T_{(:,:,)}(V^{(t-1)}, W^{(t-1)})((\Gamma(V^{(t-1)}, W^{(t-1)})+\lambda_2 I_R)^T \oplus \lambda_1 L_a)^{-1},$$

ISBN 978-93-88808-61-3

which is the least-square solution by fixing $V^{(t-1)}$ and $W^{(t-1)}$. By rearranging Eq. (20), $U^{(t)}$ can be derived as

$$\begin{aligned}
&U^{(t-1)} - \eta^{(t)}((\Gamma(V^{(t-1)}, W^{(t-1)}) + \lambda_2 I_R)^T \oplus \lambda_2 L_a)^{-1} \\
&(-\nabla_U \ell(\cdot, V^{(t-1)}, W^{(t-1)}) + U^{(t-1)}(\Gamma(V^{(t-1)}, W^{(t-1)}) + \lambda_2 I_R) \\
&+ \lambda_1 L_a U^{(t-1)}) \\
=&U^{(t-1)} + \eta^{(t)} \nabla_U \ell(\cdot, V^{(t-1)}, W^{(t-1)})((\Gamma(V^{(t-1)}, W^{(t-1)}) \\
&+ \lambda_2 I_R)^T \oplus \lambda_2 L_a)^{-1} - \eta^{(t)} U^{(t-1)} \\
=&(1 - \eta^{(t)}) U^{(t-1)} + \eta^{(t)} U^*.
\end{aligned}$$

Therefore, $U^{(t)}$ is a linear combination of $U^{(t-1)}$ and U^*. When $V^{(t-1)}$ and $W^{(t-1)}$ are fixed as constants, according to [53], U^* is a solution of the least-square problem, and thus

$$L(T, U^*, V^{(t-1)}, W^{(t-1)}) \leq L(T, U, V^{(t-1)}, W^{(t-1)}), \forall U.$$

Let U in the right-hand side be 0, and we have

$$L(T, U^*, V^{(t)}, W^{(t)}) \leq \frac{\eta^{(t)}}{2} \|T\|^2. \quad (0)$$

Therefore, since the Frobenius norm of T is bounded, the initial solution of U is also bounded, and so is $U^{(t)}$. The updates of $V^{(t)}$ and $W^{(t)}$ are similar. The theorem follows. □

Notice that this theorem does not guarantee the convergence[9] of the algorithm since the solutions may oscillate when updating U, V, and W sequentially. Nevertheless, the above property is useful in practice since the solution can be always bounded and does not result in overflow [54].

5 Experimental Results

In this section, we evaluate SNMDD with real datasets. A user study with 3126 people is conducted to evaluate the accuracy of SNMDD. Moreover, a feature study is performed. Finally, we apply SNMDD on large-scale datasets and analyze the detected SNMD types.

Data Preparation and Evaluation Plan

In the following, we detail the preparation of the datasets used in our evaluation.

User Study

We recruit 3126 OSN users around the world via Amazon Mechanical Turk (MTurk) to obtain data for training and testing the classifiers in SNMDD. The participants include 1790 males and 1336 females. Their professions are very diverse, affiliating with universities, government offices, technology companies, art centers, banks, and businesses. Each user is first invited to fill out the standard SNMD questionnaires [1], [18].[10] Then, a group of professional psychiatrists participating in this project assess and manually label the users as *potential SNMD cases* (and their types of SNMDs) or *normal users*.[11] There are 389 users labeled as *SNMD*, including 246 Cyber-Relationship (CR) Addiction, 267 Information Overload (IO), and 73 Net Compulsion (NC).[12] The result obtained by the psychiatrists serves as the ground truth for our evaluation. We also crawl the Facebook (denoted as FB_US) and Instagram (denoted as IG_US) data of the participants in the user study for training and testing of our machine learning models (based on features detailed in Section 3.1). All the data are collected with the Facebook and Instagram APIs as listed in Table 1.

TABLE 1
Details of the datasets

Dataset	Description
FB_US	User profile, the friends of each user, the news feeds created by users with metadata (who likes, who comments, stickers, and geotag), the news feeds users like orcomment (stickers also), events (join/decline), joined groups with events, and game posts created by game apps
IG_US	User profile, the followers/followees of each user, the media created by users with metadata (who likes, who comments, and geotag), and the contents users like or comment
FB_L	Anonymized user ID that performs the action, anonymized user ID that receives the action, and timestamp of action creation
IG_L	Anonymized media ID, anonymized ID of the user who created the media, timestamp of media creation, set of tags assigned to the media, number of likes and number of comments received

In the experiment, we first evaluate the effectiveness of the proposed features, including all features (All), social interaction features (Social), personal profile features (Personal), with a baseline feature Duration, i.e., the total time spent online, using TSVM [13] for semi-supervised learning in the user study. The combinations of different features, i.e., Duration with Social (D-S), Duration with Personal (D-P), Social with Personal (S-P), are also presented. We also collect two large-scale datasets, including Facebook (denoted as FB_L) with 63K nodes, 1.5M edges, and 0.84M wall posts [19], and Instagram (denoted as IG_L) with 2K users, 9M tags, 1200M likes, and 41M comments [20]. Note that some proposed features cannot be extracted from certain large-scale datasets, e.g., game posts and stickers are not available in IG_L, which is handled by using the imputation technique [21]. The details of the data crawled from each social media are listed in Table 1.

With labeled (IG_US and FB_US) and unlabeled data (IG_L and FB_L) described above, we perform a 5-fold cross validation, i.e., take 4 folds for training and 1 fold for testing, to evaluate the performance of proposed features using semi-supervised TSVM. A number of supervised approaches, including J48 Decision Tree Learning [22], SVM [23], Logistic Regression, and DTSVM [24] which do not use unlabeled data, are also compared to justify our choice of using TSVM in SNMDD. Next, we compare the proposed *SNMD-based Tensor Model (STM)*, implemented by different algorithms, i.e., *STM-Tucker-SGD*, *STM-CP-SGD*, and *STM-CP-2SGD*, with two baseline algorithms. The first baseline algorithm concatenates the features from different networks together (denoted as *CF*), while the second baseline algorithm employs the existing Tucker model (denoted as *Tucker*) that does not incorporate prior knowledge regarding the characteristics of SNMD cases (observed from our analysis).

9. To ensure the convergence of 2SGD, more assumptions, e.g., $(\alpha, \gamma, s, \delta)$-strict saddle with stochastic gradient oracle with the radius at most Q, are required to prove the range to optimal solutions [51], [52], but the above condition may not always exist in practical situations.

10. The IRB number of this project is AS-IRB-HS 15003 v.1.

11. They are from California School of Professional Psychology, Taipei City Hospital, Nat'l Taipei Univ., psychiatric clinics, etc.

12. Note that a person may have multiple types of SNMDs simultaneously.

Finally, the effectiveness of each feature is carefully analyzed in Section 5.5.

Large-Scale Experiments

To discover new insights, we apply our semi-supervised SNMDD on IG_L and FB_L to classify their users and then analyze the detected cases of different SNMD types. Notice that the goal of this analysis is exploratory-oriented as we do not have the ground truth for the large datasets. We examine whether friends of SNMD cases tend to be potential SNMD cases as well. Also, we apply community detection on FB L and IG_L to derive the relationships between different types of SNMD users in their communities. Finally, the average hop distance between the SNMD users of the same type is reported.

Evaluation of the Proposed Features

In the following, we first evaluate the performance of the proposed features using TSVM. We adopt Accu- racy (Acc.) and Area Under Curve (AUC) for evalua- tion of SNMDD. Moreover, Microaveraged-F1 (Micro-F1) and Macroaveraged-F1 (Macro-F1) are also compared for multiple-label classification. Table 2 summarizes the average results and standard deviations, where the examined feature sets are denoted by self-explained labels.

The results on the IG_US and FB_US datasets in the user study show that Durationleads to the worst performance, i.e., the results of accuracy are 34% and 36%, and the AUC are 0.362 and 0.379, respectively. Notice that the AUC function can flip the results if the calculated AUC is less than 0.5, i.e., 1-AUC. Here, we do not filp the results to show that Durationis in fact a negative predictor in our case because Duration cannot differentiate heavy users with addictive users. Using all (All) or parts (Social or Personal) of the proposed features outperforms Duration significantly (see Table 2). All achieves the best performance (80% and 84% accuracy on the IG_US and FB_US datasets, respectively) because SNMDD is able to capture the various features ex- tracted from data logs to effectively detect SNMD cases. The performance of Personal and Social are comparable, and the integrated feature set All outperforms Personal and Social by at least 15% and 16% on IG_US and FB_US in terms of accuracy. Since the F1 measure ignores true negatives, its magnitude is mostly determined by the number of true positives, i.e., large classes dominate small classes in microaveraging. As shown in Table 2, Micro-F1 of Duration, Social, and Personal are larger than Macro- F1 using both IG_US and FB_US datasets, indicating that using parts of features performs better on IO and CR (large classes) than NC. In contrast, the performance of SNMDD is almost the same in Micro-F1 and Macro-F1, which indicates its robustness. The results from FB_US are better than those from IG_US because IG_US is sparser, e.g., there are no event and game posts on Instagram. After comparing the results from SNMDD with the ground truth obtained via user study, we observe that some false-positive users are detected as NC, probably because people with NC are more likely to hide their real usage time, e.g., the game logs of some people with NC are hidden. As a result, a few normal users may be incorrectly detected as NC. However, SNMDD generally performs very well for NC due to some effective features. For example, users of NC are usually less parasocial since they are less frequent to interact with friends. Moreover, since the NC users' friends with game benefits usually do not know the NC users' other friends (e.g., colleagues), their clustering coefficients are lower than the normal users. Finally, the performance of Social and Personal with Duration features (D-S and D-P) are almost the same since SVM finds the best hyperplane to classify the training data and may not take the dimensions that downgrade the results. The results also manifests that the proposed features are robust with SVM. The p-value tests of different features with All indicate that All is significantly better than Duration, Social, Personal, D-S, D-P with p values that are much smaller than 0.05, while the performance is close to S-P.

Evaluation of Classification Techniques and STM

In the following, given all the proposed features, we first evaluate TSVM in comparison with some representative supervised learning approaches in SNMDD. As shown in Table 3, the accuracy of semi-supervised TSVM (84.3%) outperforms all the supervised algorithms, including A_2- regularized logistic regression (78.6%) and A_2-regularized A_2-loss SVM (79.2%), since TSVM effectively uses unlabeled data to address the issues of overfitting and data sparsity. The accuracy and AUCs of the single-source supervised learning methods are similar, indicating that the proposed features provide robust information that is not sensitive to the choice of learning algorithms.

Next, we compare the proposed multi-source *STM- Tucker-SGD*, *STM-CP-SGD*, and *STM-CP-2SGD* with two baselines, i.e., *CF* and *Tucker*, to integrate the features ex- tracted from IG_US and FB__US datasets with TSVM. Table 3 points out that the accuracy and AUC of *STM-CP-2SGD* are 90.4% and 0.938, respectively. *STM-CP-2SGD* with the decomposed latent factor matrix U can effectively recover important missing features and provide extra latent infor- mation to better characterize the users. In contrast, *CF*, which simply concatenates the features from FB_US and IG_US, suffers from the worst accuracy and AUC and is even beated by single-source A_2-regularized A_2-loss SVM. This is because *CF* loses correlations among the features and thereby tends to introduce noises. On the other hand, *STM-CP-2SGD* outperforms the other approaches because it incorporates important characteristics of SNMD and thereby derives more precise and accurate latent features, while the accuracy and AUC of *STM-Tucker-SGD* and *STM-CP-SGD* are almost the same.

Evaluation of the Proposed Tensor Decomposition Acceleration

In the following, the default dimensionality of U, V and W, threshold *s*, and the maximum number of iterations are set as 10, 0.001, and 50, respectively. We first compare the loss function through each iteration in Figure 2(a). Note that *STM-CP-2SGD* converges very quickly (always terminates before the 5-th iteration). Between *STM-Tucker- SGD* and *STM-CP-SGD*, the loss of *STM-Tucker-SGD* is slightly smaller since the core tensor of *STM-Tucker-SGD*

ISBN 978-93-88808-61-3

TABLE 2
Different combinations of feature categories for performance evaluations on the IG US and FB_US datasets.

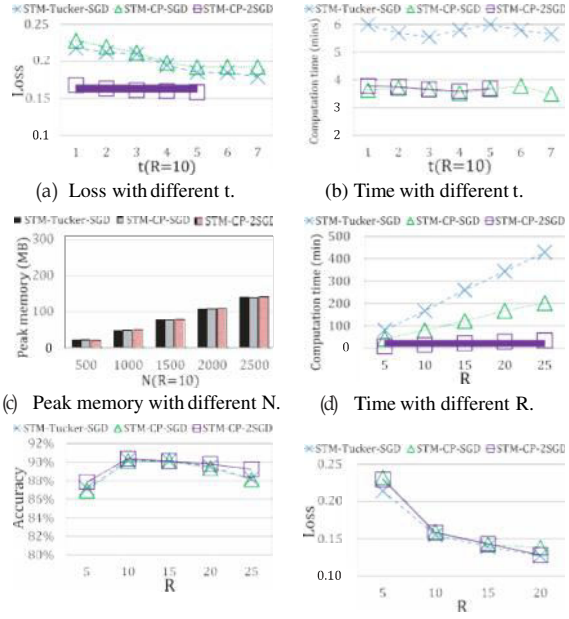
Instagram Measure	Duration	Social	Personal	D-S	D-P	S-P	All
Acc.	0.34±0.02	0.67±0.01	0.69±0.03	0.63±0.02	0.68±0.02	0.80±0.01	0.80±0.01
AUC	0.36±0.02	0.71±0.02	0.74±0.01	0.69±0.02	0.73±0.02	0.81±0.01	0.81±0.01
Micro-F1	0.33±0.01	0.76±0.01	0.78±0.04	0.74±0.01	0.77±0.03	0.85±0.01	0.85±0.01
Macro-F1	0.33±0.01	0.71±0.01	0.73±0.02	0.69±0.02	0.72±0.02	0.85±0.01	0.85±0.01
p value on AUC	$3.80 \cdot 10^{-8}$	$6.05 \cdot 10^{-5}$	$1.18 \cdot 10^{-5}$	$1.64 \cdot 10^{-6}$	$3.06 \cdot 10^{-6}$	0.76	-
Facebook							
Acc.	0.36±0.01	0.72±0.03	0.73±0.02	0.70±0.03	0.73±0.01	0.84±0.02	0.84±0.02
AUC	0.37±0.01	0.75±0.02	0.77±0.02	0.74±0.01	0.76±0.02	0.86±0.02	0.85±0.01
Micro-F1	0.44±0.04	0.80±0.02	0.81±0.01	0.79±0.02	0.80±0.01	0.90±0.01	0.90±0.01
Macro-F1	0.35±0.02	0.76±0.01	0.77±0.03	0.74±0.01	0.76±0.02	0.91±0.01	0.91±0.01
p value on AUC	$7.01 \cdot 10^{-9}$	$2.35 \cdot 10^{-5}$	$3.06 \cdot 10^{-4}$	$3.90 \cdot 10^{-6}$	$1.46 \cdot 10^{-4}$	0.064	-

TABLE 3
Comparisons of SNMDD with different classification techniques.

Technique	Acc.	AUC
Single-source (FB)		
J48 Decision Tree Learning	75.4%	0.763
A_1-regularized A_2-loss SVM	79.1%	0.790
A_2-regularized A_2-loss SVM	79.2%	0.791
A_1-regularized logistic regression	78.5%	0.788
A_2-regularized logistic regression	78.6%	0.789
DTSVM	78.5%	0.782
TSVM	**84.2%**	**0.851**
Multi-source (FB+IG)		
CF	76.4%	0.775
Tucker	87.9%	0.892
STM-Tucker-SGD	**90.2%**	**0.933**
STM-CP-SGD	**90.1%**	**0.930**
STM-CP-2SGD	**90.4%**	**0.938**

allows more freedom to fit data. On the other hand, *STM- CP-2SGD* outperforms both *STM-Tucker-SGD* and *STM-CP- SGD* in terms of the convergence rate, as well as the loss function. Figure 2(b) compares the running time through each iteration. The results manifest that the running time of *STM-CP-SGD* is the fastest for each iteration, with *STM-CP-2SGD* as the close second (which terminates first). They both significantly outperforms *STM-Tucker*. The overall running time of *STM-CP-2SGD* is the smallest since it requires much fewer iterations. Figure 2(c) shows the peak memory usage of different tensor decomposition methods. The memory usage of *STM-CP-2SGD* is slightly greater than that of the others. Notice that the feature tensor and adjacency matrix are sparse, and therefore we use sparse tensor representa- tion for each decomposition to reduce the memory usage.

Moreover, Figures 2(d)-(f) compare the performance of *STM-Tucker-SGD*, *STM-CP-SGD*, and *STM-CP-2SGD* in terms of running time, accuracy, and loss function with different *R*, respectively. As shown in Figures 2(d), *STM- CP-2SGD* significantly outperforms the other two in terms of running time for different *R*, i.e., the running time is at most 9.7% and 18.9% of *STM-Tucker-SGD* and *STM-CP-SGD*, re- spectively, while the accuracy for detecting SNMD is almost the same for different proposed methods as shown in Figure 2(e), which shows the power of acceleration of the proposed *STM-CP-2SGD* without sacrificing accuracy. Moreover, the accuracy of different methods does not increase as *R* grows since the latent features may overfit the training data and thus do not perform well on testing data. Figure 2(f) further shows the loss function with different *R*. As *R* increases, the loss functions for different methods all decrease. For



(a) Loss with different t. (b) Time with different t.
(c) Peak memory with different N. (d) Time with different R.
(e) Accuracy with different R. (f) Loss with different R.

Fig. 2. Comparisons of different datasets.

Figure 2(e) and Figure 2(f), although the loss function of *STM-Tucker-SGD* slightly outperforms *STM-CP-2SGD*, the accuracy of *STM-Tucker-SGD*, *STM-CP-SGD*, and *STM-CP- 2SGD* are similar. In summary, *STM-CP-2SGD* is the most efficient one without compromising efficiency and accuracy.

Feature Study

To observe the differences among the three types of SN- MDs, Table 4 lists the top-5 discriminative features using information gain and the corresponding accuracy on the FB_US dataset by TSVM, where CC, BI, BL, and SD re- spectively denote the clustering coefficient, burst intensity, burst length, and standard deviation. It is worth noting that the number of selfies, an indicator of self-disclosure, is not useful for detecting CR and IO, but it is effective for NC. This is because NC users are usually less socially active, comparing to CR and IO users. Moreover, the online/offline interaction ratio of NC is much higher than the ratios of the other two types, probably because NC users usually show less willingness to join offline activities. In contrast, CR

ISBN 978-93-88808-61-3

and IO cases prefer to use social media, instead of playing games alone. Moreover, people with compulsive personality are more introverted. In contrast, people with CR usually create virtual bonds to develop pathological relationships for compensation of their (missing) offline relationships. The Shannon index, an indicator of the social diversity, is also

useful in detecting NC since the friends of NC are in similar backgrounds, and thus the Shannon index is lower than that of normal users. Moreover, the social comparison score

is important for detecting NC cases. This is because when the users with malicious envy see the positive newsfeeds from the friends with similar backgrounds, they may eager

to pursue the sense of success, which is much easier to be achieved in online games.

The parasociality, effective for detecting all SNMD types, is especially useful for detecting CR cases. For example, in our user study, we find user A, 21-year-old male, frequently posting news feeds, such as "I'm so bored :((((...Ahhh- hhh!!", and his cross-dressing photos on his Facebook time-

line, more than 3 times a week, which usually get fewer than 5 likes. At the same time, he "likes" a large number of posts from others. SNMDD classifies him as a potential CR case and his questionnaire reveals that he constantly blocks out

disturbing thoughts about life and finds himself anticipating when he goes online again.

Burst intensity and length are quite useful for detecting IO cases. For example, user B, 36-year-old male, is detected as IO since the behavior of clicking "likes" fits the pattern of bursts, i.e., the median of his burst intensity is high, equal to 31. His answers to the standard questionnaire reveal that he loses sleep due to late-night access on Facebook to check others' news feeds. Through interview, user B explains that he cannot stop checking for new posts and e-mails even when all his news feeds and emails are read. Some of his friends reply him: "are you a robot? no sleep needed?!?!!", indicating that user B is indulged in finding social news. Moreover, social roles are important in detecting IO since the users with IO usually share or like the information from different communities ant thus are inclined to be detected as structural holes.

Next, we analyze the importance of different features

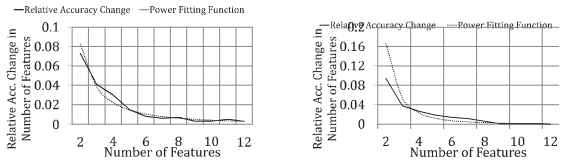
to our classifiers. χ^2-test is exploited to measure the impor- tance of each feature via SelectKBest of Scikit-Learn. The top 5 important features overall are 1) median of the intensity of bursts, 2) parasociality, 3) online/offline interaction ratio,

4) number of used stickers, and 5) standard deviation of the length of bursts. It is worth noting that TSVM using only these 5 features in SNMDD achieves an accuracy of 80.7% for FB_ US, close to that of using all features (All). In other words, integrating important social and personal features provides good results because effective personal features, e.g., the temporal behavior features, can be used to differentiate the users suffering from withdraw orrelapse symptoms and heavy users, while social features capture the interactions among users to differentiate different SNMDs.

Figs. 3(a) and 3(b) show the improvement made by adding different features in TSVM on the FB_US dataset and the proposed *STM* on multi-source data (i.e., FB_US and IG_US). The feature selection of TSVM is based on the information gain (the top-5 features mentioned earlier), while the tensor approach automatically extracts important

TABLE 4
Top features and Acc. on the FB US dataset.

CR	NC	IO	
Parasociality	Game posts	Median of BI	Median of BI
Median of BI	Online/offline ratio	Online/offline ratio	Online/offline ratio
Eigenvector centrality	Parasociality	SD of BL	Parasociality
Online/offline ratio	Shannon index	Sticker number	Sticker number
Sticker number	Social comparison score	Social roles	SD of BL
Acc.: 80.5%	Acc.: 77.6%	Acc.: 82.9%	Acc.: 80.7%



(a) Relative improvement w.r.t. the number of features.

(b) Relative improvement w.r.t. the number of latent features (*STM*).

Fig. 3. Relative accuracy change with respect to number of features.

latent features. We observe a diminishing return property on both figures, where the improvement becomes marginal as more features are included. Fig. 3(a) shows a power fit

function ($p(x) = 0.3091x^{-1.92}$) of the curve with $R^2 = 0.9534$. The exponent 1.92 denotes that the improvement by adding n-th feature is $n^{-1.92}$ times smaller than that by adding the first feature. On the other hand, the results

of the tensor-based approach in Fig. 3(b) show that the accuracy increment for adding a single feature drops faster

($p(x) = 1.11x^{-2.01}$) since the proposed *STM* can extract much more important and concise features.

Analysis of SNMD Types in Large Datasets

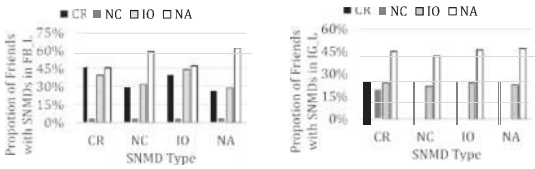
In this analysis, we first apply the proposed SNMDD frame- work (with TSVM) on some large-scale OSN datasets, i.e., FB_L and IG_L, to classify their users. In Figs. 4(a) and 4(b), we analyze the detected SNMD cases among the friends of an SNMD user. In Fig. 4(a), the leftmost bar indicates that in FB_L, among all CR users, about 45% of their friends are also CR users, which is greater than the percentage of other SNMD types. On the other hand, the 8th bar from the left in Fig. 4(a) indicates that in FB_L, about 59% of NC users' friends are NA (non-SNMD users). Figs. 4(a) and 4(b) show that, in FB_L and IG_L, CR and IO users have similar friend types. This is because CR and IO cases, by their nature, are similar, i.e., they are both seeking social satisfaction (e.g., relationships and information) from the OSNs. Moreover, among different SNMD cases, CR and IO users are likely to be friends with other CR and IO users. For CR users, this phenomenon has been described as "loneliness propagates" [15].

Furthermore, Infomap community detection [39] is per- formed on FB_L and IG_L to derive the relationships be- tween different types of SNMD users in their communities. Figs. 4(c) and 4(d) analyze the community structures of

ISBN 978-93-88808-61-3

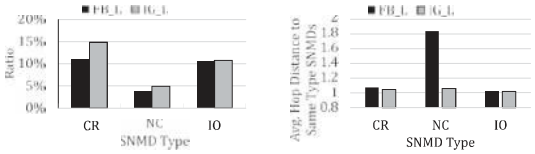
SNMD users with different SNMD scores, where each point represents the characteristic of a community. Specifically, each community in the dataset is represented by three differ- ent types of points, i.e., CR, NC, and IO. For example, each CR point is represented as *(score, ratio)*, where *score* is the average CR score in that community, and *ratio* indicates the proportion of CR users in the community. It is similar for each IO/NC point. As Figs. 4(c) and 4(d) show, for each SNMD type, when the average SNMD score is higher, it is likely to have more SNMD users in the community. Moreover, there are many communities with large IO scores in IG_L that have IO ratios close to 1. This implies that the users with large IO scores in IG_L are more inclined to form homogeneous groups. At the first glance, one may feel that NC users frequently appear in many communities, and there seems to be a large number of NC users, especially in FB_L (i.e., Fig. 4(c)). However, after carefully examining these communities, we find that those communities (with large ratios of NC users) are usually very small (usually with the size around 5) because NC users are less-active. On the other hand, in IG_L, when SNMD scores are larger, the ratios of IO users in communities are also larger. This is because IO users can view, like, or follow others in Instagram more easily (not necessary to be friends first).

Fig. 4(e) compares the ratios of different types of SNMD users identified in FB_L and IG_L. There are more CR users in IG_L probably because CR users seek social supports online to compensate the loneliness in real life. We argue that the Instagram platform makes it easy to freely create social relationships with strangers. In contrast, it is not that easy to create new social relationships on Facebook since the friend requests need to be approved. Finally, Fig. 4(f) compares the average number of hops from each SNMD user to the nearest user with the same type of SNMDs. The leftmost bar shows that the average hop distance from each CR user to the closest CR user is 1.07 hop, indicating that CR and IO users are close to other same-type users, i.e., average hop distances are within 1.15, where Figs. 4(a) and 4(b) also report similar results.



(a) SNMD types of friends (FB_L). (b) SNMD types of friends (IG_L).

(c) Ratios of SNMD users in communities (FB_L). (d) Ratios of SNMD users in communities (IG_L).



(e) SNMD users. (f) Hop distance to same type SNMD users.

Fig. 4. Comparisons of different datasets.

6 Conclusion

In this paper, we make an attempt to automatically identify potential online users with SNMDs. We propose an SNMDD framework that explores various features from data logs of OSNs and a new tensor technique for deriving latent fea- tures from multiple OSNs for SNMD detection. This work represents a collaborative effort between computer scien- tists and mental healthcare researchers to address emerging issues in SNMDs. As for the next step, we plan to study the features extracted from multimedia contents by techniques on NLP and computer vision. We also plan to further explore new issues from the perspective of a social network service provider, e.g., Facebook or Instagram, to improve the well-beings of OSN users without compromising the user engagement.

References

[1] K. Young, M. Pistner, J. O'Mara, and J. Buchanan. Cyber-disorders: The mental health concern for the new millennium. *Cyberpsychol. Behav.*, 1999.

[2] J. Block. Issues of DSM-V: internet addiction. *American Journal of Psychiatry*, 2008.

[3] K. Young. Internet addiction: the emergence of a new clinical disorder, *Cyberpsychol. Behav.*, 1998.

[4] I.-H. Lin, C.-H. Ko, Y.-P. Chang, T.-L. Liu, P.-W. Wang, H.-C. Lin, M.-F. Huang, Y.-C. Yeh, W.-J. Chou, and C.-F. Yen. The associa- tion between suicidality and Internet addiction and activities in Taiwanese adolescents. *Compr. Psychiat.*, 2014.

[5] Y. Baek, Y. Bae, and H. Jang. Social and parasocial relationships on social network sites and their differential relationships with users' psychological well-being. *Cyberpsychol. Behav. Soc. Netw.*, 2013.

[6] D. La Barbera, F. La Paglia, and R. Valsavoia. Social network and addiction. *Cyberpsychol. Behav.*, 2009.

[7] K. Chak and L. Leung. Shyness and locus of control as predictors of internet addiction and internet use. *Cyberpsychol. Behav.*, 2004.

[8] K. Caballero and R. Akella. Dynamically modeling patients health state from electronic medical records: a time series approach. *KDD*, 2016.

[9] L. Zhao and J. Ye and F. Chen and C.-T. Lu and N. Ramakr- ishnan. Hierarchical Incomplete multi-source feature learning for Spatiotemporal Event Forecasting. *KDD*, 2016.

[10] E. Baumer, P. Adams, V. Khovanskaya, T. Liao, M. Smith, V. Sosik, and K. Williams. Limiting, leaving, and (re)lapsing: an exploration of Facebook non-use practices and experiences. *CHI*, 2013.

[11] R. Jain and N. Abouzakhar. A comparative study of hidden markov model and support vector machine in anomaly intrusion detection. *JITST*, 2013.

[12] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. *KDD*, 2011.

[13] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Large scale trans- ductive svms. *JMLR*, 2006.

[14] L. Leung. Net-generation attributes and seductive properties of the internet as predictors of online activities and internet addic- tion. *Cyberpsychol. Behav. Soc. Netw.*, 2004.

[15] J. Cacioppo, J. Fowler, and N. Christakis. Alone in the crowd: the structure and spread of loneliness in a large social network. *J. Pers. Soc. Psychol.*, 2009.

[16] J. Kleinberg. Bursty and hierarchical structure in streams. *KDD*, 2002.

[17] K.-L. Liu, W.-J. Li, and M. Guo. Emoticon smoothed language models for twitter sentiment analysis. *AAAI*, 2012.

[18] C. Andreassen, T. Torsheim, G. Brunborg, and S. Pallesen. Devel- opment of a Facebook addiction scale. *Psychol. Rep.*, 2012.

[19] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in Facebook. *WOSN*, 2009.

[20] E. Ferrara, R. Interdonato, and A. Tagarelli. Online popularity and topical interests through the lens of Instagram. *HT*, 2014.

[21] M. Saar-Tsechansky and F. Provost. Handling missing values when applying classification models. *JMLR*, 2007.
[22] I. Witten and E. Frank. Data mining: practical machine learn- ing tools and techniques with Java implementations. Morgan- Kaufmann, San Francisco, 2000.
[23] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001.
[24] F. Chang, C.-Y. Guo, X.-R. Lin, and C.-J. Lu. Tree decomposition for large-scale SVM problems. *JLMR*, 2010.
[25] P. Comon, X. Luciani, and A. L. D. Almeida. Tensor decomposi- tions, alternating least squares and other tales. *Journal of Chemo- metrics*, 2009.
[26] E. Acar, D. M. Dunlavy, and T. G. Kolda. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics*, 2011.
[27] L. He, C.-T. Lu, J. Ma, J. Cao, L. Shen, and P. S. Yu. Joint community and structural hole spanner detection via harmonic modularity. *KDD*, 2016.
[28] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummad. Measur- ing user influence on twitter: The million follower fallacy. *ICWSM*, 2010.
[29] K. Hayashi, T. Maehara, M. Toyoda, and K. Kawarabayashi. Real- time top-r topic detection on twitter with topic hijack filtering. *KDD*, 2015.
[30] J. Gill and G. King. What to do when your Hessian is not invert- ible: Alternatives to model respecification in nonlinear estimation. *Sociological Methods and Research*, 2004.
[31] B. Kågström and P. Poromaa. Distributed and shared memory block algorithms for the triangular Sylvester equation with sep^{-1} estimators. *SIAM Journal on Matrix Analysis and Applications*, 1992.
[32] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 1966.
[33] R.A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers in Phonetics*, 1970.
[34] S. Graham, A. Munniksma, J. Juvonen. Psychosocial benefits of cross-ethnic friendships in urban middle schools. *Child Develop- ment*, 2013.
[35] S. S. Levine, E. P. Apfelbaum, M. Bernard, V. L. Bartelt, E. J. Zajac, and D. Stark. Ethnic diversity deflates price bubbles. *National Academy of Sciences*, 2014.
[36] H. Appel, J. Crusius, and Alexander L. Gerla. Social comparison, envy, and depression on facebook: a study looking at the effects of high comparison standards on depressed individuals. *Journal of Social and Clinical Psychology*, 2015.
[37] A. Bordes, L. Bottou, and P. Gallinari. SGD-QN: careful quasi- newton stochastic gradient descent. *Journal of Machine Learning Research*, 2009.
[38] J. B. White, E. J. Langer, L. Yariv, and J. C. Welch IV. Frequent social comparisons and destructive emotions and behaviors: the dark side of social comparisons. *Journal of Adult Development*, 2006.
[39] M. Rosvall and C. Bergstrom. Maps of random walks on complex networks reveal community structure. *Natl. Acad. Sci.*, 2008.
[40] D. L. King, P. H. Delfabbro, D. Kaptsis, and T. Zwaans. Adolescent simulated gambling via digital and social media: an emerging problem. *Computers in Human Behavior*, 2014.
[41] D. Li, X. Li, L. Zhao, Y. Zhou, W. Sun, and Y. Wang. Linking multiple risk exposure profiles with adolescent Internet addiction: insights from the person-centered approach. *Computers in Human Behavior*, 2017.
[42] K. Kim, H. Lee, J. P. Hong, M. J. Cho, M. Fava, D. Mischoulon, D. J. Kim, and H. J. Jeon. Poor sleep quality and suicide attemptamong adults with internet addiction: a nationwide community sample of Korea. *PLOS ONE*, 2017.
[43] C.-H Chang, E. Saravia, and Y.-S. Chen. Subsonscious crowd- sourcing: a feasible data collection mechanism for mental disorder detection on social media, *ASONAM*, 2016.
[44] B. Saha, T. Nguyen, D. Phung, and S. Venkatesh. A framework for classifying online mental health-related communities with an interest in depression. *IEEE Journal of Biomedical and Health Informatics*, 2016.
[45] M. Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. *ICWSM*, 2013.
[46] T. Kolda and B. Bader. Tensor decompositions and applications. *SIAM review*, 2009.
[47] H.-H. Shuai, C.-Y. Shen, D.-N. Yang, Y.-F. Lan, W.-C. Lee, P. S. Yu, and M.-S. Chen. Mining online social data for detecting social network mental disorders. *WWW*, 2016.
[48] A. Anandkumar, D. Hsu, M. Janzamin, and S. Kakade. When are overcomplete topic models identifiable? uniqueness of tensor tucker decompositions with structured sparsity. *JLMR*, 2015.
[49] L. Bottou. Stochastic gradient descent tricks. *Neural Networks: Tricks of the Trade*, 2012.
[50] J. M. Ortega and W. C. Rheinboldt. Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, NY, 1970.
[51] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 2006.
[52] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. *Confer- ence of Learning Theory (COLT)*, 2015.
[53] T. Maehara, K. Hayashi, and K. Kawarabayashi. Expected tensor decomposition with stochastic gradient descent. *AAAI*, 2016.
[54] V. Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 2006.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Nearest Keyword Set Search in Multi-Dimensional Datasets

Srinivas Reddy,
M.Tsech
Scholar, CSE
Department,
Malla Reddy College of Engineering

Abstract—Keyword-based search in text-rich multi-dimensional datasets facilitates many novel applications and tools. In this paper, we consider objects that are tagged with keywords and are embedded in a vector space. For these datasets, we study queries that ask for the tightest groups of points satisfying a given set of keywords. We propose a novel method called ProMiSH (Projection and Multi Scale Hashing) that uses random projection and hash-based index structures, and achieves high scalability and speedup. We present an exact and an approximate version of the algorithm. Our experimental results on real and synthetic datasets show that ProMiSH has up to 60 times of speedup over state-of-the-art tree-based techniques.

Index Terms—Querying, multi-dimensional data, indexing, hashing

1 INTRODUCTION

OBJECTS (e.g., images, chemical compounds, documents, or experts in collaborative networks) are often charac- terized by a collection of relevant features, and are com- monly represented as points in a multi- dimensional feature space. For example, images are represented using color fea- ture vectors, and usually have descriptive text information (e.g., tags or keywords) associated with them. In this paper, we consider multi-dimensional datasets where each data point has a set of keywords. The presence of keywords in feature space allows for the development of new tools to query and explore these multi-dimensional datasets.

In this paper, we study *nearest keyword set* (referred to as NKS) queries on text-rich multi-dimensional datasets. An NKS query is a set of user-provided keywords, and the result of the query may include k sets of data points each of which contains all the query keywords and forms one of the top-k tightest cluster in the multi- dimensional space. Fig. 1 illustrates an NKS query over a set of two-dimensional data points. Each point is tagged with a set of keywords. For a query Q ¼ **f***a; b; c*g, the set of points **f**7*;* 8*;* 9g contains all the query keywords **f***a; b; c*g and forms the tightest cluster com- pared with any other set of points covering all the query keywords.

Therefore, the set f7*;* 8*;* 9g is the top-1 result for the query Q.

NKS queries are useful for many applications, such as photo-sharing in social networks, graph pattern search, geo- location search in GIS systems[1] [1], [2], and so on. The fol- lowing are a few examples.

1) Consider a photo-sharing social network (e.g.,Facebook), where photos are tagged with people names and locations. These photos can be embed- ded in a high-dimensional feature space of texture, color, or shape [3], [4]. Here an NKS query can find a group of similar photos which contains a set of people.
2) NKS queries are useful for graph pattern search, where labeled graphs are embedded in a high dimensional space (e.g., through Lipschitz embed- ding [5]) for scalability. In this case, a search for a subgraph with a set of specified labels can be answered by an NKS query in the embedded space [6].
3) NKS queries can also reveal geographic patterns. GIS can characterize a region by a high-dimensional set of attributes, such as pressure, humidity, and soil types. Meanwhile, these regions can also be tagged with information such as diseases. An epidemiolo- gist can formulate NKS queries to discover patterns by finding a set of similar regions with all the dis- eases of her interest.

We formally define NKS queries as follows.

Nearest keyword set. Let Dc R^d be a d-dimensional data-set with N points. For any o 2 D, it is tagged with a set of keywords $s\eth o\flat = \mathbf{f}v_1; :::; v_t\mathrm{g} \leq V$, where V is a dictionary of U unique keywords. For any $o_i; o_j$ 2 D, the distance between o_i and o_j is measured by their L_2-norm (i.e., euclidean distance) as $dist\eth o_i; o_j\flat$ ¼ ko_i — o_jk$_2$. Given a set

ISBN 978-93-88808-61-3

of data points A c D, rðA Þ is the diameter A and is defined by the maximum distance between any two

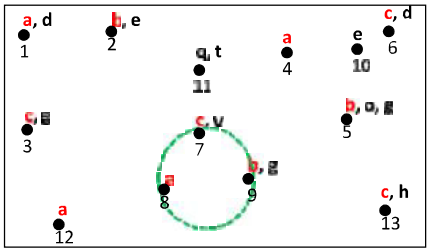


Fig. 1. An example of an NKS query on a keyword tagged multi-dimen- sional dataset. The top-1 result for query $\mathbf{f}a; b; c\mathbf{g}$ is the set of points f7*;* 8*;* 9g.

in Q by $Q \leq$ S sðvÞ. Let S be the set including all candi- dates of Q. The top-1 result A^m of Q is obtained by

$$A^m \text{¼} \arg\min r(A)$$

:

$A2S$

Similarly, a top-k NKS query retrieves the top-k candidates with the least diameter. If two candidates have equal diame- ters, then they are further ranked by their cardinality.

Although existing techniques using tree-based indexes [2], [7], [8], [9] suggest possible solutions to NKS queries on multi-dimensional datasets, the performance of these algorithms deteriorates sharply with the increase of size or dimensional- ity in datasets. Our empirical results show that these algo- rithms may take hours to terminate for a multi-dimensional dataset of millions of points. Therefore, there is a need for an efficient algorithm that scales with dataset dimension, and yields practical query efficiency on large datasets.

In this paper, we propose ProMiSH (short for Projection and Multi-Scale Hashing) to enable fast processing for NKS queries. In particular, we develop an exact ProMiSH (referred to as ProMiSH-E) that always retrieves the opti- mal top-k results, and an approximate ProMiSH (referred to as ProMiSH-A) that is more efficient in terms of time and space, and is able to obtain near-optimal results in practice. ProMiSH-E uses a set of hashtables and inverted indexes to perform a localized search. The hashing tech- nique is inspired by Locality Sensitive Hashing (LSH) [10], which is a state-of-the-art method for nearest neighbor search in high-dimensional spaces. Unlike LSH-based methods that allow only approximate search with probabi- listic guarantees, the index structure in ProMiSH-E sup- ports accurate search. ProMiSH-E creates hashtables at multiple bin-widths, called index levels. A single round of search in a hashtable yields subsets of points that contain query results, and ProMiSH-E explores each subset using a fast pruning-based algorithm. ProMiSH-A is an approxi- mate variation of ProMiSH-E for better time and space effi- ciency. We evaluate the performance of ProMiSH on both real and synthetic datasets and employ state-of-the-art VbR^m-Tree [2] and CoSKQ [8] as baselines. The empirical results reveal that ProMiSH consistently outperforms the baseline algorithms with up to 60 times of speedup, and ProMiSH-A is up to 16 times faster than ProMiSH-E obtaining near-optimal results.

Our main contributions are summarized as follows. (1) We propose a novel multi-scale index for exact and approximate NKS query processing. (2) We develop effi- cient search algorithms that work with the multi-scale indexes for fast query processing. (3) We conduct exten- sive experimental studies to demonstrate the performance of the proposed techniques.

The paper is organized as follows. We start with the related work in Section 2. Next, we present the index struc- ture for exact search (ProMiSH-E) in Section 3, the exact search algorithm in Section 4, and its optimization techni- ques in Section 5. In addition, we introduce the approximate algorithm (ProMiSH-A) and provide an analysis for its approximation ratio in Section 6. The time and space com- plexity for ProMiSH are analyzed in Section 7. Experimental results are presented in Section 8. Finally, We conclude this paper with future work in Section 9. A glossary of the nota- tions is shown in Table 1.

2 Related Work

A variety of related queries have been studied in literature on text-rich spatial datasets.

Location-specific keyword queries on the web and in the GIS systems [11], [12], [13], [14] were earlier answered using a combination of R-Tree [15] and inverted index. Felipe et al. [16] developed IR^2-Tree to rank objects from spatial datasets based on a combination of their distances to the query locations and the relevance of their text descriptions to the query keywords. Cong et al. [17] integrated R-tree and inverted file to answer a query similar to Felipe et al. [16] using a different ranking function. Martins et al. [18] computed text relevancy and location proximity independently, and then combined the two ranking scores.

Cao et al. [7] and Long et al. [8] proposed algorithms to retrieve a group of spatial web objects such that the group's keywords cover the query's keywords and the objects in the group are nearest to the query location and have the lowest inter-object distances. Other related queries include aggre- gate nearest keyword search in spatial databases [19], top-k preferential query [20], top-k sites in a spatial data based on their influence on feature points [21], and optimal location queries [22], [23].

Our work is different from these techniques. First, existing works mainly focus on the type of queries

ISBN 978-93-88808-61-3

the coordinates of query points are known [7], [8]. Even though it is possible to make their cost functions same to the cost function in NKS queries, such tuning does not change their techniques. The proposed techniques use location information as an integral part to perform a best- first search on the IR-Tree, and query coordinates play a fundamental role in almost every step of the algorithms to prune the search space. Moreover, these techniques do not provide concrete guidelines on how to enable efficient processing for the type of queries where query coordi- nates are missing. Second, in multi- dimensional spaces, it is difficult for users to provide meaningful coordinates, and our work deals with another type of queries where users can only provide keywords as input. Without query coordinates, it is difficult to adapt existing techniques to our problem. Note that a simple reduction that treats the coordinates of each data point as possible query coordi- nates suffers poor scalability. Third, we develop a novel index structure based on random projection with hashing. Unlike tree-like indexes adopted in existing works, our index is less sensitive to the increase of dimensions and scales well with multi-dimensional data.

Another track of related works deal with *m*-closest keywords queries [9]. In [9], bR*-Tree is developed based on a R*-tree [24] that stores bitmaps and minimum bounding rectangles (MBRs) of keywords in every node along with points MBRs. The candidates are generated by the *apriori* algorithm [25]. Unwanted candidates are pruned based on the distances between MBRs of points or keywords and the best found diameter. However, the pruning techniques become ineffective with an increase in the dataset dimension as there is a large overlap between MBRs due to the curse of dimensionality. This leads to an exponential number of candidates and large query times. A poor estimation of starting diameter further worsens the performance of their algorithm. bR*-Tree also suffers from a high storage cost; therefore, Zhang et al. modified bR*-Tree to create Virtual bR*- Tree [2] in memory at run time. Virtual bR*-Tree is created from a pre-stored R*- Tree, which indexes all the points, and an inverted index which stores keyword information and path from the root node in R*-Tree for each point. Both bR*-Tree and Virtual bR*-Tree, are structurally similar, and use similar candi- date generation and pruning techniques. Therefore, Virtual bR*-Tree shares similar performance weaknesses as bR*-Tree.

Tree-based indexes, such as R-Tree [15] and M-Tree [26], have been extensively investigated for near- est neighbor search in high-dimensional spaces. These indexes fail to scale to dimensions greater than 10 because of the curse of dimensionality [27]. Random projection [28] with hashing [10], [29], [30], [31], [32] has come to be the state-of- the-art method for nearest neighbor search in high- dimensional datasets. Datar et al. [10] used random vectors constructed from *p*-stable distributions to project points, computed hash keys for the points by splitting the line of projected values into disjoint bins, and then concatenated hash keys obtained for a point from *m* ran- dom vectors to create a final hash key for the point. Our

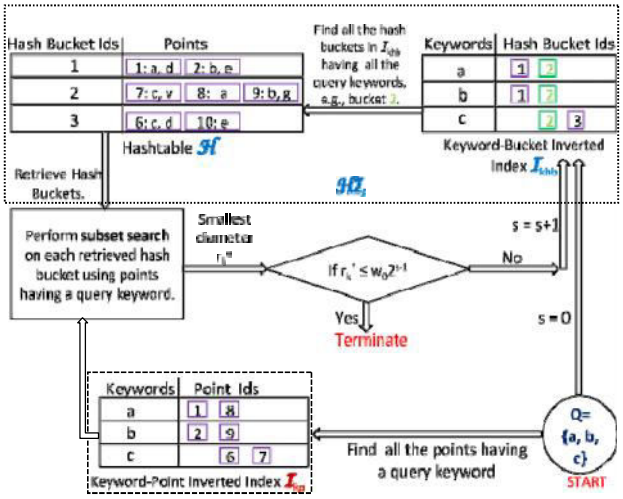


Fig. 2. Index structure and flow of execution of ProMiSH.

find the top-*k* tightest clusters that cover the input keyword set. Meanwhile, nearest neighbor queries usually require coordinate information for queries, which makes it difficult to develop an efficient method to solve NKS queries by existing techniques for nearest neighbor search. In addition, multi-way distance joins for a set of multi-dimensional datasets have been studied in [33], [34]. Tree-based index is adopted, but suffers poor scalability with respect to the dimension of the dataset. Furthermore, it is not straightforward to adapt these algorithms since every query requires a multi-way distance join only on a subset of the points of each dataset.

3 Index Structure for Exact ProMiSH

We start with the index for exact ProMiSH (ProMiSH-E). This index consists of two main components.

Inverted Index $\mathbf{I}_{kp}$. The first component is an inverted index referred to as $\mathbf{I}_{kp}$. In $\mathbf{I}_{kp}$, we treat keywords as keys, and each keyword points to a set of data points that are associated with the keyword. Let D be a set of data points and V be a dictionary that contains all the keywords appear- ing in D. We build $\mathbf{I}_{kp}$ for D as follows. (1) For each $v \in V$, we create a key entry in $\mathbf{I}_{kp}$, and this key entry points to a set of data points $D_v = \{o \in D \mid v \in \text{doc}(o)\}$ (i.e., a set includes all data points in D that contain keyword *v*). (2) We repeat

(1) until all the keywords in V are processed. In Fig. 2, an example for $\mathbf{I}_{kp}$ is shown in the dashed rectangle at the bottom.

Hashtable-Inverted Index Pairs HI. The second component consists of multiple hashtables and inverted indexes referred to as HI. HI is controlled by three parameters: (1) (*Index level*) *L*, (2) (*Number of random unit vectors*) *m*, and (3) (*hashtable size*) *B*. All the three parameters are non-negative integers. Next, we describe how these three parameters con- trol the construction of HI.

In general, HI contains *L* hashtable-inverted index pairs, characterized by $\{(H^{0,s}, \mathbf{I}^{0,s}) \mid s \in \{0, 1, 2, \ldots, L-$

ISBN 978-93-88808-61-3

problem is different from nearest neighbor search. NKS queries provide no coordinate information, and aim to

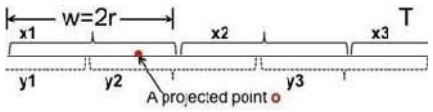


Fig. 3. The projection space (a segment) of a random unit vector is parti- tioned into overlapping bins of equal width: point *o* is included in bin *x*1 and *y*2.

First, given a set of *d*-dimensional data points D, we cre- ate hashtable $H^{\vartheta_{s}Þ}$ as follows.

1) We randomly sample *m d*-dimensional unit vectors $z_1; z_2; \ldots ; z_m$ (i.e., $kz_ik_2 ¼ 1$ for $i ¼ 1; 2; \ldots ; m$);
2) For each *o* 2 D, we compute its projection on each of the unit vectors $o_{z_i} ¼ o \square z_i$ where $i ¼ 1; 2; \ldots ; m$;
3) Let *pMax* be the maximum projected value for data points in D. For each z_i, we consider its projection space as a segment $[0; pMax]$, and partition the segment into $2^{\vartheta L-s Þ þ} 1$ overlapping bins, where each bin has width $w ¼ w_0 2^s$ and is equally overlapped with two other bins as shown in Fig. 3. We conduct the projection space partition on all the *m* random unit vectors.
4) For each z_i and *o* 2 D, since its projection space is partitioned into overlapping bins, o_{z_i} falls into two bins; therefore, we get two bin ids $\{b_1(o; z_i); b_2(o; z_i)\}$, and we can compute $b_1(o; z_i)$ and $b_2(o; z_i)$ as below,

$$b_1(o; z_i) = \left\lfloor \frac{o_{z_i} - w/2}{w} \right\rfloor \quad (1)$$

$$b_2(o; z_i) = \begin{cases} b_1(o; z_i) + 1, & \text{if } o_{z_i} - \left\lfloor \frac{o_{z_i} - w/2}{w} \right\rfloor w \geq \frac{3w}{4} \\ b_1(o; z_i) - 1, & \text{otherwise.} \end{cases} \quad (2)$$

Fig. 3 shows an example where we partition the proj- ect space into overlapping bins $\{x1; x2; x3; y1; y2; y3\}$, and point *o* lies in bins *x*1 and *y*2.
5) For each *o* 2 D, we generate its signatures based on the bins into which its projections on random unit vectors fall. With *m* random unit vectors, we obtain *m* pairs of bin ids for each data point *o*. Next, we take a cartesian product over these *m* pairs of bin ids and generate 2^m signatures for each point *o*, where each signature $\{b_{i1}(o; z_1); \ldots ; b_{im}(o; z_m)\}$ contains a bin id from each of the *m* pairs. For example, let z_1 and z_2 be two random unit vectors, and the bin ids for a point *o* be $\{x1; y1\}$ from z_1 and $\{x2; y2\}$ from z_2. We create four signatures as $\{x1; x2\}$, $\{x1; y2\}$, $\{y1; x2\}$, and $\{y1; y2\}$.
6) For each point *o* 2 D, we hash it into 2^m buckets in $H^{\vartheta_{s}Þ}$ using its 2^m signatures. For each signature $\{b_{i1}(o; z_1); \ldots ; b_{im}(o; z_m)\}$, we convert it into a hashtable bucket id by a standard hash function, $\left(\sum_{j=1}^{m} b_{ij} \cdot pr_j\right) \% B$, where *B* is hashtable size (i.e., the

to

where $H^{\vartheta_{s}Þ}$ and $I^{\vartheta_{s}Þ}$ are the *s*-th hashtable and inverted index, respectively.

each of which contains at least one data point *o* such that *v* 2 $s(o)$. Fig. 2 demonstrates an example about HI with one pair of hashtable and inverted index shown in the dotted rectangle.

In next section, we show how to conduct exact search using ProMiSH-E.

4 Search Algorithm for ProMiSH-E

In this section, we present the search algorithms in ProM- iSH-E that finds top-*k* results for NKS queries. First, we introduce two lemmas that guarantee ProMiSH-E always retrieves the optimal top-*k* results. Then, we describe the details in ProMiSH-E. We start with some theoretic results for ProMiSH-E.

Lemma 1. *Let* R^d *be a d-dimensional euclidean space and z be a random unit vector uniformly sampled from* R^d *such that* $kzk_2 ¼ 1$. *For any two points* o_1 2 R *and* o_2 2R , *we have* $ko_1 — o_2k_2 \leq ko_{1z} — o_{2z}k_2$, *where* o_{1z} *and* o_{2z} *are the projec- tion of* o^1 *and o* ²*on z, respectively.*

Proof. Since Euclidean space with dot product is an inner product space, we have

$$ko_{1z} — o_{2z}k_2 ¼ jð o_1 — o_2 Þ_z j \leq kzk \square k\, o\, — o\, k$$
$$¼ ko_1 — o_2k_2; \text{ since } kzk_2 ¼ 1:$$

The inequality follows the Cauchy-Schwarz inequality. ■

Lemma 2. *Given* $A ¼ \{o ; \ldots ; o\ g$ c R^d *with diameter r is pro- jected onto a d-dimensional random unit vector z and the pro- jection space of z is partitioned into overlapping bins with equal width w, there exists at least one bin containing all the points in A if* $w \leq 2r$.

Proof. According to Lemma 1 and the definition of diameter, we have $8o_i; o_j 2 A; jo_{iz} — o_{jz}j \leq ko_i — o_jk \leq r_2$ Thus, we can further derive $\max(o_{1z}; \ldots ; o_{nz}) — \min(o_{1z}; \ldots ; o_{nz}) \leq r$. Since the projection space of *z* is partitioned into overlap- ping bins of width $w \leq 2r$, it follows from the construction that any line segment of width *r* is fully contained in one of the bins as shown in Fig. 3. Hence, all the points in *A* will fall into the same bin. ■

We use an example to show how Lemma 2 guarantees the retrieval of the optimal top-1 results. Given a query *Q*, we assume the diameter of its top-1 result is *r*. We project all the data points in D on a unit random vector and parti- tion the projected values into overlapping bins of bin-width $w \leq 2r$. If we perform a search in each of the bins indepen- dently, then Lemma 2 guarantees that the top-1 result of query *Q* will be found in one of the bins. Based on Lemma 1 and Lemma 2, we propose ProMiSH-E as shown in Fig. 2. A search starts with the HI structure

ISBN 978-93-88808-61-3

number of buckets in $H^{\vartheta,P}$) and pr_j is a random prime number.

Second, given a dictionary V and hashtable $H^{\vartheta,P}$, we cre- ate the inverted index $\mathbf{I}^{\vartheta,\flat}$. In this inverted index, keys are still keywords. For each v 2 V_{khb} v points to a set of buckets,

at index level s ¼ 0. ProMiSH-E finds the buckets in hasht- able $H^{\vartheta 0,P}$, each of which contains all the query keywords by inverted index $\mathbf{I}^{\vartheta 0,\flat}$. Then, ProMiSH-E explores each selected bucket using an efficient pruning based technique to generate results. ProMiSH-E terminates after exploring

HI structure at the smallest index level s such that all the top-k results have been found.

---

the bucket from the hashtable H, and filters these points using bitset BS to get a subset of points F^0 in steps (17-22).
Subset F^0 contains only those points which are taggedwith

Algorithm 1. ProMiSH-E

In: Q: query keywords; k: number of top results
In: w_0: initial bin-width
1: PQ ½e½]; þ1Þ]: priority queue of top-k results 2: HC: hashtable to check duplicate candidates
3: BS : bitset to track points having a query keyword 4: for all o 2 [$_{8v_Q 2Q}$ I$_{kp}$½v_Q] do
5: BS½o] true /* Find points having query keywords*/ 6: end for
7: for all s 2 f0; … ;L — 1g do
8: Get HI at s
9: E½] 0 /* List of hash buckets */ 10: for all v_Q 2 Q do 11: for all bld 2 I$_{khb}$½v_Q] do
12: E½bld] E½bld] þ 1
13: end for
14: end for
15: for all i 2 ð0; … ; $SizeOf$ðEÞÞ do
16: if E½i] ¼ $SizeOf$ðQÞ then
17: F^0 ; /* Obtain a subset of points */ 18: for all o 2 H½i] do
19: if BS½o] ¼ true then
20: F^0 F^0 [o
21: end if
22: end for
23: if checkDuplicateCand(F^0; HC) = falsethen 24: searchInSubset(F^0, PQ)
25: end if
26: end if
27: end for
28: /* Check termination condition */ 29: if PQ½k]:$r \leq w$ 2^{s-1} then
30: Return PQ
31: end if
32: end for
33: /* Perform search on D if algorithm has not terminated */ 34: for all o 2 D do
35: if BS½o] ¼ true then
36: F^0 F^0 [o

at least one query keyword and is explored further.

Subset F^0 is checked whether it has been explored earlier or not using *checkDuplicateCand* (Algorithm 2) in step 23. Since each point is hashed using 2^m signatures, duplicate subsets may be generated. If F^0 has not been explored earlier, then ProMiSH-E performs a search on it using *searchInSubset* (Algorithm 3) at step 24 (We discuss in this algorithm in detail in Section 5). Results are inserted into a priority queue PQ of size k. Each entry of PQ is a tuple containing a set of points and their diame- ter. PQ is initialized with k entries, each of whose set is empty and the diameter is þ1. Entries of PQ are ordered by their diameters, and entries with equal diameters are further ordered by their set sizes. A new result is inserted into PQ only if its diameter is smaller than the k-th small- est diameter in PQ. If ProMiSH-E does not terminate after exploring the HI structure at index level s, then the search proceeds to HI at index level ðs þ 1Þ.

Algorithm 2. CheckDuplicateCand

In: F^0: a subset; HC: hashtable of subsets 1: F^0 $sort$ðF^{0}Þ
2: $pr1$: list of prime numbers; $pr2$: list of primenumbers; 3: for all o 2 F^0 do
4: pr_1 randomSelectð$pr1$Þ; pr_2 randomSelectð$pr2$Þ 5: h_1 h_1 þ ðo × pr_1Þ; h_2 h_2 þ ðo × pr_2Þ
6: end for
7: h $h_1 h_2$;
8: if isEmptyðHC½h]Þ=false then
9: if elementWiseMatch(F^0; HC½h]) = true then
10: Return true;
11: end if
12: end if
13: HC½h].addðF^0Þ;
14: Return false;

ProMiSH-E terminates when the k-th smallest diameter

Algorithm 1 details the steps in ProMiSH-E. It maintains a bitset BS. For each v_Q 2 Q, ProMiSH-E retrieves the list of points corresponding to v_Q from I$_{kp}$ in step 4. For each point o in the retrieved list, ProMiSH-E marks the bit correspond- ing to o's identifier in BS as true in step 5. Thus, ProMiSH-E finds all the points in D which are tagged with at least one query keyword. Next, the search continues in the HI struc- tures, beginning at s ¼ 0. For any given index level s, ProMiSH-E works with H$^{ðs\text{Þ}}$ and I$^{ðs\text{Þ}}_{khb}$ in HI at step 8. ProMiSH-E retrieves all the lists of hash bucket ids corresponding to keywords in Q from the inverted index I$^{ð0\text{Þ}}_{kp}$ at steps (10-11). An intersection of these lists yields a set of hash buckets each of which contains all the query keywords in steps (12-16) (e.g., In Fig. 2, this intersection yields the bucket id 2). For each selected hash bucket, ProMiSH-E retrieves all the points in

ISBN 978-93-88808-61-3

Lemma 2 guarantees that all the possible candidates are fully contained in one of the bins of the hashtable, and therefore, must have been explored. If ProMiSH-E fails to terminate after exploring HI at all the index levels $s \in \{0, \ldots, L-1\}$, then it performs a search in the complete dataset D during steps(34-39).

Algorithm *checkDuplicateCand* (Algorithm 2) uses a hashtable *HC* to check duplicates for a subset F'. Points in F' are sorted by their identifiers. Two separate standard hash functions are applied to the identifiers of the points in the sorted order to generate two hash values in steps (2-6). Both of the hash values are concatenated to get a hash key *h* for the subset F' in step 7. The use of multiple hash func- tions helps to reduce hash collisions. If *HC* already has alist of subsets at *h*, then an element-wise match of F' is per- formed with each subset in the list in steps (8-9). Otherwise, F' is stored in *HC* using key *h* in step 13.

As shown in Algorithm 1, the efficiency of ProMiSH- E highly depends on an efficient search algorithm that finds top-*k* results from a subset of data points. In next sec- tion, we propose a search algorithm that provides such efficiency.

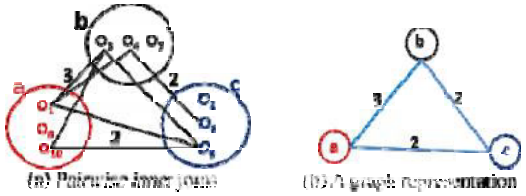


Fig. 4. (a) *a*, *b*, and *c* are groups of points of a subset F' obtained for a query $Q=\{a; b; c\}$. A point *o* in a group *g* is joined to a point o' in another group g' if $\|o - o'\| \leq r_k$. The groups in the order $\{a, c, b\}$ generates the least number of candidates by a multi-way join. (b) A graph of pairwise inner joins. Each group is a node in the graph. The weight of an edge is the num- ber of point pairs obtained by an inner join of the corresponding groups.

5 Search in a Subset of Data Points

We present an algorithm for finding top-*k* clusters tightest random. Let the edge *ac*, with weight 2, be selected in Fig. 4b. This forms the ordered set $\{a — c\}$. The next edge to be selected is the least weight edge such that at least one of its nodes is not included in the ordered set. Edge *cb*, with weight 2, is pickednextin Fig. 4b. Now the orderedsetis $\{a — c — b\}$. This process terminates when all the nodes are included in the set. $\{a — c — b\}$ gives the orderingof the groups.

Algorithm 3 shows how the groups are ordered. The *k*th smallest diameter r_k is retrieved form the priority queue *PQ* in step 1. For a given subset F' and a query *Q*, all the points are grouped using query keywords in steps (2-5). A pair- wise inner join of the groups is performed in steps (6-18). An adjacency list *AL* stores the distance between points which satisfy the distance threshold r_k. An adjacency list *M* stores the count of point pairs obtained for each pair of groups by the inner join. A greedy algorithm finds the order of the groups in steps (19-30). It repeatedly removes an edge with the smallest weight from *M* till all the groups are included in the order set *curOrder*. Finally, groups are sorted using *curOrder* in step 30. in a subset of points. A subset is obtained from a hasht-

We explain a multi-way distance join with an example. A multi-way distance join of *q* groups $\{g_1; \ldots; g_q\}$ finds all the tuples $\{o_{1;i}; \ldots; o_{x;j}; o_{y;k}; \ldots; o_{q;l}\}$ such that $\forall x, y$: $o_{x;j} \in g_x$, $o_{y;k} \in g_y$, and $\|o_{x;j} — o_{y;k}\|_2 \leq r_k$. Fig. 4a shows groups $\{a, b, c\}$ of points obtained for a query $Q = \{a, b, c\}$ from a subset F'. We show an edge between a pair of points of two groups if the distance between the points is at most r_k, e.g., an edge between point o_1 in group *a* and point o_3 in group *b*. A multi- way distance join of these groups finds tuples $\{o_1, o_3, o_9\}$ and $\{o_{10}, o_3, o_9\}$. Each tuple obtained by a multi-way join is a promising candidate for a query.

Group Ordering

A suitable ordering ofthe groupsleads to anefficient candi- date exploration by a multi-way distance join. We first per- form a pairwise inner joins of the groups with distance threshold r_k. In inner join, a pair of points from two groups are joined only if the distance between them is at most r_k. Fig. 4a shows such a pairwise inner joins of the groups $\{a, b, c\}$. We see from Fig. 4a that a multi-way distance join in the order $\{a, b, c\}$ explores 2 true candidates $\{\{o_1, o_3, o_9\}, \{o_{10}, o_3, o_9\}\}$ and a false candidate $\{o_1, o_4, o_6\}$. A multi-way distance join in the order $\{a, c, b\}$ explores the least number of candidates 2. Therefore, a proper ordering of the groups leads to an effective pruning of false candidates. Optimal orderingofgroupsfortheleast numberofcandidates gener- ation is NP-hard [35].

We propose a greedy approach to find the ordering of groups. We explain the algorithm with a graph in Fig. 4b. Groups $\{a, b, c\}$ are nodes in the graph. The weight of an edge is the count of point pairs obtained by an inner join of the corresponding groups. The greedy method starts by selecting an edge having the least weight. If there are multiple edges with the same weight, then an edge is s lected at

Algorithm 3. SearchInSubset

In: F': subset of points; *Q*: query keywords; *q*: query size
In: *PQ*: priority queue of top-*k* results
1: r_k *PQ*.*k*]:*r* /* *k*th smallest diameter */
2: *SL* ½ð*v*;½]Þ]: list of lists to store groups per query keyword

ISBN 978-93-88808-61-3

```
for all v 2 Q do
    SL[v] ← {o ∈ F[l] : o is tagged with v} /* form groups */
end for
/* Pairwise inner joins of the groups*/
AL: adjacency list to store distances between points
M: adjacency list to store count of pairs between groups
for all (v_i, v_j) ∈ Q such that i ≤ q, j ≤ q, i < j do
    for all o ∈ SL[v_i] do
        for all o' ∈ SL[v_j] do
            if ||o − o^0||_2 ≤ r_k then
                AL[o, o'] ← ||o − o'||_2
                M[v_i, v_j] ← M[v_i, v_j] + 1
            end if
        end for
    end for
end for
/* Order groups by a greedy approach */ curOrder = []
while Q ≠ ∅ do
    (v_i, v_j) ← removeSmallestEdge(M)
    if v_i ∉ curOrder then
        curOrder.append(v_i); Q ← Q − v_i
    end if
    if v_j ∉ curOrder then
        curOrder.append(v_j); Q ← Q − v_j
    end if
end while
sort(SL, curOrder) /* order groups */
findCandidates(q, AL, PQ, Idx, SL, curSet, curSetr, r_k)
```

Nested Loops with Pruning

We perform a multi-way distance join of the groups by nested loops. For example, consider the set of points in

Fig. 4. Each point $o_{a;i}$ of group a is checked against each point $o_{b;j}$ of group b for the distance predicate, i.e., $\|o_{a;i} - o_{b;j}\|_2 \leq r_k$. If a pair $(o_{a;i}, o_{b;j})$ satisfies the distance predicate, then it forms a tuple of size 2. Next, this tuple is checked against each point of group c. If a point $o_{c;k}$ satisfies the distance predicate with both the points $o_{a;i}$ and $o_{b;j}$, then a tuple $(o_{a;i}, o_{b;j}, o_{c;k})$ of size 3 is generated. Each intermedi- ate tuple generated by nested loops satisfies the property that the distance between every pair of its points is at most r_k. This property effectively prunes false tuples very early in the join process and helps to gain high efficiency. A can- didate is found when a tuple of size q is generated. If a can- didate having a diameter smaller than the current value of r_k is found, then the priority queue PQ and the value of r_k are updated. The new value of r_k is used as distance thresh- old for future iterations of nestedloops.

Algorithm 4. findCandidates

```
In: q: query size; SL: list of groups
In: AL: adjacency list of distances between points
In: PQ: priority queue of top-k results
In: Idx: group index in SL
In: curSet: an intermediate tuple
In: curSetr: an intermediate tuple's diameter 1:          if Idx ≤ q then
 2:     for all o 2 SL[Idx] do
 3:       if AL[curSet[Idx-1], o] ≤ r_k then
 4:         newCurSetr curSetr
 5:         for all o' 2 curSetdo
 6:           dist    AL[o, o']
 7:           if dist ≤ r_kthen
 8:             flag  true
 9:             if newCurSetr ≺ dist then
10:               newCurSetr dist
11:             end if
12:           else
13:             flag false; break;
14:           end if
15:         end for
16:         if flag = true then
17:           newCurSet  curSet.append(o)
18:           r_k findCandidates(q, AL, PQ, Idx þ 1; SL, newCurSet, newCurSetr, r_k)
19:         else
20:           Continue;
21:         end if
22:       end if
23:     end for
24:     return r_k
25:   else
26:     if checkDuplicateAnswers(curSet, PQ) = true then
27:       return r_k
28:     else
```

An intermediate tuple *curSet* is checked against each point of group $SL[Idx]$ in steps (2-23). First, it is determined using AL whether the distance between the last point in *curSet* and a point o in $SL[Idx]$ is at most r_k in step 3. Then, the point o is checked against each point in *curSet* for the dis- tance predicate in steps (5-15). The diameter of *curSet* is updated in steps (9-11). If a point o satisfies the distance predicate with each point of *curSet*, then a new tuple *newCurSet* is formed in step 17 by appending o to *curSet*. Next, a recursive call is made to *findCandidates* on the next group $SL[Idx + 1]$ with *newCurSet* and *newCurSetr*. A can- didate is found if *curSet* has a point from every group. A result is inserted into PQ after checking for duplicates in steps (26-33). A duplicate check is done by a sequential match with the results in PQ. For a large value of k, a method similar to Algorithm 2 can be used for a duplicate check. If a new result gets inserted into PQ, then the value of r_k is updated in step 18.

6 Approximate Search: ProMiSH-A

In this section, we discuss the approximate version of ProM- iSH referred to as ProMiSH-A. We start with the algorithm description of ProMiSH-A, and then analyze its approxima- tion quality.

Algorithm overview. In general, ProMiSH-A is more time and space efficient than ProMiSH-E, and is able to obtain near-optimal results in practice. The index structure and the search method of ProMiSH-A are similar to ProMiSH-E; therefore, we only describe the differences between them.

The index structure of ProMiSH-A differs from ProM- iSH-E in the way of partitioning projection space of random unit

ISBN 978-93-88808-61-3

vectors. ProMiSH-A partitions projection space into non-overlapping bins of equal width, unlike ProMiSH-E which partitions projection space into overlapping bins. Therefore, each data point *o* gets one bin id from a random unit vector *z* in ProMiSH-A. Only one signature is generated for each point *o* by the concatenation of its bin ids obtained from each of the *m* random unit vectors. Each point is hashed into a hashtable using its signature.

The search algorithm in ProMiSH-A differs from ProM- iSH-E in the termination condition. ProMiSH-A checks for a termination condition after fully exploring a hashtable at a given index level: It terminates if it has *k* entries with non- empty data point sets in its priority queue *PQ*.

Approximation quality analysis. In the following, we ana- lyze the approximation quality for the top-1 result returned by ProMiSH-A. In particular, we use approximation ratio r ≤ 1 as the metric to evaluate approximation quality. This ratio is defined as the ratio of the diameter of the result reported by ProMiSH-A *r* to the diameter of the optimal result r^m: r ¼ $r=r^m$. Let D be a *d*-dimensional dataset and

29: if *curSetr* ≺ *PQ*[*k*]:*r* then 30: *PQ.Insert*([*curSet*, *curSetr*]) 31:
32: end if
33: end if
34: end if
return *PQ*[*k*]:*r*

We find results by nested loops as shown in Algorithm 4

that each data point in D has *t* keywords, and each keyword is independently sampled by a uniform distribution over a dictionary V with *U* unique keywords. We define $f(v) = 1 - (1 - 1/U)^t$ as the probability that a data point has keyword *v* 2 V. Thus, we can estimate the expected number of points that have keyword *v* as $E[N_v] = Nf(v)$. To this end, the expected number of candidates for query *Q* in D is estimated by N_Q

(*findCandidates*). Nested loops are performed recursively.

Let $g(r)$ be the probability that a candidate has a diameter no more than *r*. Then, the expected number of candidates for query *Q* with diameter no more than *r* is estimated by

$$N_r = g(r) N^q \prod_{i=1}^{q} f(v_i).$$

We index data points in D by ProMiSH-A, where each data point is projected onto *m* random unit vectors. The pro- jection space of each random unit vector is partitioned into non-overlapping bins of equal width *w*. Let $Pr(A; r \mid w)$ be the conditional probability for random unit vectors that a candidate *A* of query *Q* having diameter *r* is fully contained within a bin with width *w*. For *m* independent unitrandom vectors, the joint probability that a candidate *A* is contained in a bin in each of the *m* vectors is $Pr(A; r \mid m)^m$, and the probability that no candidate of diameter *r* is retrieved by ProMiSH-A from the hashtable, created using *m* unit ran- dom vectors, is $(1 - Pr(A; r \mid w)^m)^{N_r}$. Let the diameter of the top-1 result of query *Q* be r^*. Then, the probability $P(r^*)$ of at least one candidate of any diameter *r*, where $r^* \le r \le r^k$, being retrieved by ProMiSH-A is given by

$$P(r^*) = 1 - \prod_{r=r^*}^{r^k} (1 - Pr(A; r \mid w)^m)^{N_r} : \quad (3)$$

For a given constant *Z*, $0 \le Z \le 1$, we can compute the smallest value of r^k using Equation (3) such that $Z \le P(r^*)$. The value r^k / r^* gives the approximation ratio of the results returned by ProMiSH-A with the probability *Z*.

We empirically computed r for queries of three key- words for different values of *Z* using this model. We used a 32-dimensional real dataset having one million points described in Section 8 for our investigation, and computed the values of N_r and $Pr(A; r \mid w)^2$, where we use two random unit vectors with bin-width of $w = 100$. In this way, we obtained the approximation ratio bound of r ¼ 1:4 and r ¼ 1:5 for *Z* ¼ 0:8 and *Z* ¼ 0:95, respectively.

in a bucket *b* among *B* buckets. Suppose ProMiSH-E applies *m* random unit vectors. Since ProMiSH-E generates 2^m sig- natures for each data point, the expectation of $N_{b;v}$ under uniformity assumptions is estimated as below,

$$E[N_{b;v}] = \frac{2^m}{B} E[N_v]:$$

Searching a bucket *b* in $H^{(s)}$ includes inner group joins and nested loops. Let *q* be the number of keywords in a query. First, inner group joins for *d*-dimensional data points are computed in $O(d(E[N_{b;v}])^2 + q(E[N_{b;v}])^2)$. Second, nested loops are computed in $O((E[N_{b;v}])^q)$. Thus, the total complexity of searching a bucket *b* is $O(d(E[N_{b;v}])^2 + (qE[N_{b;v}])^2 + (E[N_{b;v}])^q)$. In the worst case, we may need to check all the buckets at all scales; therefore, the overall complexity is $O(dLB(E[N_{b;v}])^2 + LB(qE[N_{b;v}])^2 + LB(E[N_{b;v}])^q)$.

Time complexity of ProMiSH-A. Let *L* be the index level applied in the index structure of ProMiSH-A, $H^{(s)}$ be the hashtable at scale $s \in \{0; 1; \ldots; L-1\}$, *B* be hashtable size, and $N^b_{b;v}$ be the number of data points with keyword *v* lying in a bucket *b*. Since ProMiSH-A only generates one signature per data point, the expectation of $N^b_{b;v}$ under uniformity assumptions is estimated as $\frac{E[N_v]}{B}$

7 Complexity Analysis of ProMiSH

In this section, we first analyze the query time complexity and index space complexity in ProMiSH. Then we discuss how ProMiSH prunes the search space.

ISBN 978-93-88808-61-3

Query Time Complexity

Given a set of *d*-dimensional data points D, we assume data points are uniformly distributed in the buckets of a hasht- able, and keywords of each data point are uniformly sam- pled from the dictionary.

Suppose D has *N* data points, each data point has *t* key- words, and the keywords are sampled from a dictionary of *U* unique keywords. Let N_v be the number of data points with keyword *v*. The expectation of N_v is computed as follows,

$$E[N_v] = \frac{N}{B}.$$

Similarly, we can derive the overall complexity of ProMiSH- A is $O(dLB(E[N_{b,v}])^2 + LB(qE[N_{b,v}])^2 + LB(E[N_b])^2)$.

Index Space Complexity

Let *N* be the number of data points to index, *d* be the dimen- sion of data points, *t* be the number of keywords per data point, *U* be dictionary size, *m* be the number of random unit vectors for point projection, *L* be index level, and *B* be hashtable size.

Space complexity of ProMiSH-E. The Indexes of ProMiSH- E includes keyword-point inverted index $\mathbf{I}_{kp}$, hashtable H, and keyword-bucket inverted index $\mathbf{I}_{khb}$. First, we need $O(tN + U)$ space for $\mathbf{I}_{kp}$. Second, in the worst case, we need to include $O(2^m NL + BL)$ points in H. Finally, $\mathbf{I}_{khb}$ takes $O(UBL)$ space. Thus, at index level *L*, the overall space com- plexity is $O(tN + 2^m NL + UBL)$.

Space complexity of ProMiSH-A. The Indexes of ProMiSH-A also includes keyword-point inverted index $\mathbf{I}_{kp}$, hasht- able H, and keyword-bucket inverted index $\mathbf{I}_{khb}$. Unlike ProMiSH-E, H in ProMiSH-A takes at most $O(NL + BL)$. Thus, at index level *L*, the overall space complexity is $O(tN + NL + UBL)$.

Pruning Intuition

Let D be a *d*-dimensional dataset of *N* data points, *U* be dic-

ISBN 978-93-88808-61-3

Time complexity of ProMiSH-E. Let L be the index level applied in the index structure of ProMiSH-E, $H^{\vartheta s\Þ}$ be the hashtable at scale s 2 f0; 1; ... ;L — 1g, B be hashtable size, and $N_{b;v}$ be the number of data points with keyword $v$lying assume each data point is associated with only one keyword.

Suppose node set A^{m} c D with diameter r^{m} is the top-1 result for query Q. Let $f\vartheta v\Þ$ denote the probability that a

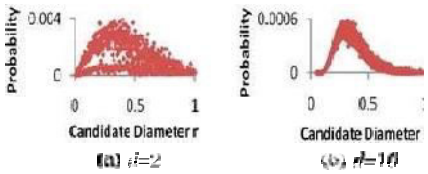


Fig. 5. Candidate diameter distributions for queries with three keywords over a two-dimensional and a 16-dimensional real datasets.

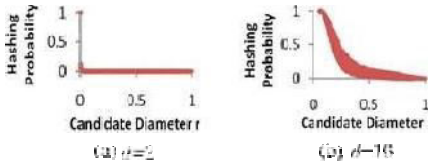


Fig. 6. Distributions of $Pr\vartheta A; rjw\Þ^{2}$ for candidates of queries with three keywords over a two-dimensional and a 16-dimensional real datasets.

data point has keyword v and $g\vartheta r\Þ$ denote the probability that a candidate of Q has diameter no more than r. Given query Q, the expected number of candidates N_Q and the expected number of candidates $N_{Q;r}$ with diameter no more than r are calculated as follows,

$$N_Q \¼ \prod_{i\¼1}^{Y} \vartheta f\vartheta v_i\ÞN\Þ; \quad N_{Q;r} \¼ g\vartheta r\ÞN_Q:$$

We select all the points in D which contain at least one query keyword v_i, project these points on a random unit vector, and split the line of projected values into overlap- ping bins of equal width $w \¼ 2r^{m}$. Let $Pr\vartheta A; r$ j $w\Þ$ be the con- ditional probability for random unit vectors that a candidate A with diameter r is fully contained within a bin of width w. For m independent random unit vectors, the probability that a candidate A is contained in a bin in each of the m vectors is $Pr\vartheta A; r$ j $w\Þ^{m}$. Ideally, the expected num- ber of candidates explored by ProMiSH in a hashtable is

$$N_p \¼ \sum_{i\¼0}^{\infty} Pr\vartheta A; \overset{m}{r} \text{ j } w_i\Þ \quad N_Q:$$

We empirically measured keyword distribution $f\vartheta\ Þ$, $Pr\vartheta A; r$ j $w\Þ^{m}$, and the ratio of N_p to N_Q by real datasets of one million data points with varied dimensions (more details about the dataset are described in Section 8).

Candidate diameter distributions and the distributions of $Pr\vartheta A; r$ j $w\Þ^{2}$ are demonstrated in Figs. 5 and 6, where candi- date diameters are scaled to between 0 and 1. We make following observations. (1) Candidate diameters follow a heavy-tailed distribution, which suggests a large number of candidates have diameters much larger than r^{m}. (2) The distributions of $Pr\vartheta A; r$ j $w\Þ^{2}$ decreases exponentially with can- didate diameter, which implies that the candidates with diameter larger than r^{m} have much smaller chance of falling in a bin and being probed by ProMiSH, compared with A^{m}. Therefore, most of candidates with diameters larger than r^{m} are effectively pruned out by ProMiSH using its index.

Table 2 presents the ratios of N_p to N_Q. Each ratio is com- puted as an average of 50 random queries. We observe that

TABLE 2
Ratios of the Expected Number of Candidates N_p to the Expected Number of Candidates N_Q

Dataset Dimension d	2	4	8	16	32
Percentage ratio ($^{Np}_{NQ}$)	0.007	0.3	5.8	22	47

TABLE 3
Statistics of Datasets Used in Experiments

ISBN 978-93-88808-61-3

Dataset	Dataset size N	Dictionary size U	Keywords per point t
Real-1	10;000	5;661	12
Real-2	30;000	6;753	13
Real-3	50;000	7;101	13
Real-4	70;000	7;902	14
Real-5	1;000;000	24;874	11

ProMiSH prunes more than 99 and 50 percent of false candi- dates for d ¼ 2 and d ¼ 32, respectively.

8 Experimental Results

In this section, we evaluate the effectiveness and efficiency of ProMiSH by both real and synthetic data.

Setup

Dataset. Our evaluation employs real and synthetic datasets. The real datasets are collected from photo-sharing web- sites. As discussed in Section 1, one of the applications in NKS queries is to find tight clusters of photos which contain all the keywords provided by a user in a photo-sharing social network. We crawl images with descriptive tags from Flickr,[2] and then these images are transformed into gray- scale. Let d be the desired dimensionality. We convert each image into a d- dimensional point by extracting its color his- togram, and associate each data point with a set of key- words that are derived from its tags. In total, we collect five datasets (referred to as Real-1, Real-2, Real-3, Real-4, and Real-5) with up to one million data points. Their statistics are shown in Table 3.

We also generate synthetic datasets to evaluate the scal- ability of ProMiSH. In particular, the data generation pro- cess is governed by the following parameters: (1) Dimension d specifies the dimensionality of each data point; (2) Dataset size N indicates the total number of multi- dimensional points in a synthetic dataset; (3) Keywords per point t suggests the number of keywords for each data point; and (4) Dictionary size U denotes the total number of keywords in a dataset. For each data point, its coordinate in each dimension is randomly sampled between 0 and 10;000, and its keyword is randomly selected following a uniform distribution. We create multiple synthetic datasets to inves- tigate how these parameters affect the performance of ProMiSH.

Query. We generate NKS queries for real and synthetic datasets. In general, the query generation process is con- trolled by two parameters: (1) Keywords per query q decides the number of keywords in each query; and (2)

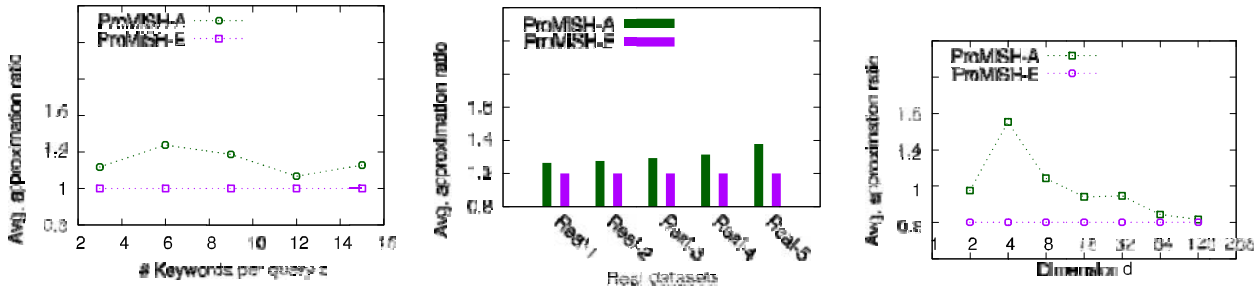


Fig. 7. Average approximation ratio of ProMiSH-A under different input real data: (left) varying the number of keywords per query q; (middle) varying real datasets; and (right) varying the number of dimensions in data points d.

Dictionary size U indicates the total number of keywords in a target dataset. For a real dataset, the probability that a key- word will be sampled in a query is proportional to the key- word's frequency in the dataset. For a synthetic dataset, a keyword of a query is randomly sampled following a uni- form distribution.

Implementation. In addition to the exact ProMiSH (ProM- iSH-E) and the approximate ProMiSH (ProMiSH-A), we also implement Virtual bR*-Tree (VbR^m-Tree) [2] and CoSKQ [7], [8] as baselines.

For VbR^m-Tree, we fix the leaf node size to be 1;000 entries and other node size to be 100 entries, as it demon- strate the best performance under this parameter setting.

CoSKQ is designed to handle the type of queries with query coordinates. To adapt CoSKQ to our problem, we transform an NKS query into a set of CoSKQ queries. Given a data point from a target dataset and an NKS query, we build a CoSKQ query by using the coordinates of the data point and the keywords in the NKS query. To ensure the correctness, if a dataset has N data points, we have to build N CoSKQ queries that enumerate all possible query coordinates.

All the algorithms are implemented in C++ with GCC 4.8.2, and all the experiments are conducted on a server with Ubuntu 14.04, powered by an Intel Core i7-2620M 2.7GHz CPU and 64 GB of RAM. Each experiment is

ISBN 978-93-88808-61-3

repeated 10 times, and their average results from 100 queries are presented.

Effectiveness

We apply real datasets to demonstrate the effectiveness of ProMiSH-A. In particular, we use the metric *approximation ratio* [30], [32] for evaluation. Let Q be an NKS query, r_i bethe ith smallest diameter from the top-k results returned by ProMiSH-A, and r^*_i be the ith smallest diameter returned by ProMiSH-E. The approximation ratio of ProMiSH-A with respect to Q is defined by $r(Q) = \frac{1}{k}\sum_{i=1}^{k} \frac{r_i}{r^*_i}$. It is easy to see $r(Q) \geq 1$; moreover, the smaller $r(Q)$ is, the better the algo- rithm will be with respect to Q. In the following, we report the average approximation ratio (AAR), which is the mean of the approximation ratios over all evaluated queries.

Fig. 7 shows the effectiveness of ProMiSH-A under dif- ferent input real data. In this set of experiments, the index parameters are fixed as $m = 4$, $L = 5$, and $B = 10{,}000$. We range the dataset among Real-1, Real-2, Real-3, Real-4, and Real-5 with Real-3 as the default dataset, the number of keywords per query q from 3 to 15 with 9 as the default q, and the number of dimensions for data points d from 2 to 128 with 16 as the default d. All the algorithms focus on top- 1 result. The results demonstrate that the AAR of ProMiSH- A is no more than 1.6 in all circumstances, and is no more than 1.2 in most cases.

Fig. 8 reports the effectiveness of ProMiSH-A under different index parameters over the real dataset Real-3. In this set of experiments, we fix the dimensions of the data points in Real-3 to be 16, and the number of keywords in queries to be 9. For index parameters, we vary the num- ber of random unit vectors m from 2 to 6 with 4 as the default m, index level L from 5 to 13 with 5 as the default m, and hashtable size B from 1,000 to 100,000 with 10,000 as the default B. All the algorithms focus on top-1 result. In general, the AAR of ProMiSH-A is no more than 1.3 in all circumstances. Moreover, we observe the following trends: (1) when m increases, the AAR increases; (2) when L increases, the AAR decreases; and (3) when B increases, the AAR increases.

Fig. 8. Average approximation ratio of ProMiSH-A under different index parameters over Real-3: (left) varying the number of random

ISBN 978-93-88808-61-3

Fig. 9 reports the effectiveness of ProMiSH-A in different top-k search over the real dataset Real-3. In this experiment, the input parameters are fixed as d ¼ 16 and q ¼ 9; and the index parameters are fixed as m ¼ 4, L ¼ 5, and B ¼ 10;000. As k is varied from 1 to 9, the AAR of ProMiSH-A is no more than 1:2.

In summary, Figs. 7, 8, and 9 consistently suggest the high effectiveness of ProMiSH-A.

Efficiency

We employ response time as the metric to evaluate the effi- ciency of different algorithms. Given a set of queries, the response time of an algorithm is defined as the average amount of time that the algorithm spends in processing one query.

Fig. 10 presents the response time of ProMiSH-E, ProM- iSH-A, and VbR^m-Tree under different input real data. In this set of experiments, the index parameters are fixed as m ¼ 4, L ¼ 5, and B ¼ 10;000. We range the dataset among Real- 1, Real-2, Real-3, Real-4, and Real-5 with Real-3 as the default dataset, the number of keywords per query q from 3 to 15 with 9 as the default q, and the number of dimensions for data points d from 2 to 128 with 16 as the default d. All the algorithms focus on top-1 result. Note that the result of CoSKQ based method is not shown in Fig. 10, as it cannot finish this experiment within one day. We make the follow- ing observations based the results. (1) As the number of key- words per query q increases, the response time of all algorithms increases. Compared with VbR^m-Tree, ProMiSH- E and ProMiSH-A are up to 30 and 60 times faster, respec- tively. (2) In all real datasets, ProMiSH-E and ProMiSH-A consistently outperform VbR^m-Tree with up to 18 and 25 times of speedup, respectively. Moreover, VbR^m-Tree can- not process the workload for Real-5 within one day. (3) When d is ranged from 2 to 128, ProMiSH-E and ProMiSH- A can finish the computation within one second. As one has to transform an NKS query into thousands of CoSKQ queries for the correctness and evaluate them all, CoSKQ based method processes a query in 2 to 10 seconds (not shown) even when d is 2 or 4, which is up to 100 times slower than our methods. In terms of VbR^m-Tree, it finishes the computation in more than one minute for d ¼ 32 (not shown), but cannot finish this experiment within one day. (4) ProMiSH-A outperforms ProMiSH-E with up to 16 times of speedup.

Fig. 11 shows the efficiency of ProMiSH-E and ProM- iSH-A under different index parameters over the real dataset Real-3. In this set of experiments, we fix the dimensions of the data points in Real-3 to be 16, and the number of keywords in queries to be 9. For index param- eters, we vary the number of random unit vectors m from 2 to 6 with 4 as the default m, index level L from 5 to 13 with 5 as the default m, and hashtable size B from 1;000 to 100;000 with 10;000 as the default B. All the algorithms focus on top-1 result. From the results, we observe that (1) m ¼ 4 empirically renders the best response time for both ProMiSH-E and ProMiSH-A; (2) as L increases, the response time of both algorithms decreases; and (3) hasht- able size has minor influence on the response time of the two algorithms.

Fig. 12 reports the response time of the algorithms on searching top-k results over Real-3. In this experiment, the input parameters are fixed as d ¼ 16 and q ¼ 9; and the index parameters are fixed as m ¼ 4, L ¼ 5, and B ¼ 10;000. Note that VbR^m-Tree and the CoSKQ based method are excluded from this experiment since they mainly support top-1 search. The results indicate that (1) as k increases, the response time of both algorithms increases; and (2) ProMiSH-A is consistently faster than ProMiSH-E.

Fig. 13 presents the response time of the algorithms under different synthetic data. In this set of experiments, the index parameters are fixed as m ¼ 4, L ¼ 5, and B ¼ 10 000, and we apply 6 parameters to control syn- thetic data generation: (1) the number of keywords per query q, ranging from 3 to 15 with 9 as the default q; (2) dataset size N, ranging from 10;000 to 10;000;000 with 1;000;000 as the default N; (3) data point dimension d, ranging from 2 to 128 with 16 as the default d; (4) the number of keywords per data point t, ranging from 1 to 16 with 4 as the default t; (5) dictionary size U, ranging from 100 to 10;000 with 1;000 as the default U; and (6) the k in top-k search, ranging from 1 to 9, with 1 as the default k. Note that the results of VbR^m-Tree and the CoSKQ based method are not shown here since they can- not finish this experiment within one day. We draw the following observations based on the results. (1) As q, N, d, t, or k increases, the response time of ProMiSH-E and ProMiSH-A increases. (2) As U increases, the response time of both algorithms decreases. (3) ProMiSH-A can process one query in 2 minutes in all cases, while ProM- iSH-E processes one query in 20 minutes in average. (4) ProMiSH-A outperforms ProMiSH-E in terms of response time with up to 100 times of speedup.

Fig. 14 reports the response time of ProMiSH-E and ProMiSH-A under different index parameters over syn- thetic data. In this set of experiments, the synthetic data are generated with a parameter setting q ¼ 9, N ¼ 1;000;000, d ¼ 16, t ¼ 4, U ¼ 1;000, and k ¼ 1. For index parameters, we range the number of random unit vectors m from 2 to 6 with 4 as the default m, index level L from 5 to 13 with 5 as the default L, and hashtable size B from 1;000 to 100;000 with 10;000 as the default B. We observe that (1) m ¼ 4 ren- ders the best response time for ProMiSH-E; (2) as L increases, ProMiSH-A obtains significant improvement in terms of efficiency; (3) hashtable size B has minor influence on both algorithms'

response time; and (4) ProMiSH-A is up to 200 times faster than ProMiSH-E.

Index Efficiency

We use memory usage and indexing time as the metrics to evaluate the index size for ProMiSH-E and ProMiSH-A. In particular, Indexing time indicates the amount of time used to build ProMiSH variants.

Fig. 15 presents the memory usage and indexing time of ProMiSH-E and ProMiSH-A under different input real

data. In this set of experiments, the index parameters are fixed as m ¼ 4, L ¼ 5, and B ¼ 10;000. We vary the num- ber of dimensions in data points m from 2 to 128 with 16 as the default m, and the datasets among Real-1, Real-2, Real-3, Real-4, and Real-5 with Real-3 as the default data- set. From the results, we make the following observations.

(1) Memory usage grows slowly in both ProMiSH-E and ProMiSH-A when the number of dimensions in data points increases. (2) ProMiSH-A is more efficient than ProMiSH-E in terms of memory usage and indexing time: it takes 80 percent less memory and 90 percent less time, and is able to obtain near-optimal results as shown in Fig. 7. (3) Over all cases, the memory usage ratio of ProM- iSH to raw data is no more than 13:4 for ProMiSH-E and no more than 2:4 for ProMiSH-A.

Summary

We summarize the experimental results as follows. First, ProMiSH-E and ProMiSH-A consistently outperform the baseline methods in terms of efficiency with up to 60 times of speedup. Second, ProMiSH-A is up to 16 times faster than ProMiSH-E, and can obtain near-optimal results. Third, ProMiSH-A is more space-efficient: compared with ProM- iSH-E, it takes 80 percent less memory and 90 percent less indexing time.

9 Conclusions and Future Work

In this paper, we proposed solutions to the problem of top-k nearest keyword set search n multi-dimensional datasets. We proposed a novel index called ProMiSH based on ran- dom projections and hashing. Based on this index, we developed ProMiSH-E that finds an optimal subset of points and ProMiSH-A that searches near-optimal results with bet- ter efficiency. Our empirical results show that ProMiSH is faster than state-of-the-art tree-based techniques, with mul- tiple orders of magnitude performance improvement. More- over, our techniques scale well with both real and synthetic datasets.

Ranking functions. In the future, we plan to explore other scoring schemes for ranking the result sets. In one scheme, we may assign weights to the keywords of a point by using techniques like tf-idf. Then, each group of points can be scored based on distance between points and weights of keywords. Furthermore, the criteria of a result containing all the keywords can be relaxed to generate results having only a subset of the query keywords.

Disk extension. We plan to explore the extension of ProM- iSH to disk. ProMiSH-E sequentially reads only required buckets from $\mathbf{I}_{kp}$ to find points containing at least one query keyword. Therefore, $\mathbf{I}_{kp}$ can be stored on disk using a direc- tory-file structure. We can create a directory for $\mathbf{I}_{kp}$. Each bucket of $\mathbf{I}_{kp}$ will be stored in a separate file named after its key in the directory. Moreover, ProMiSH-E sequentially probes HI data structures starting at the smallest scale to generate the candidate point ids for the subset search, andit reads only required buckets from the hashtable and the inverted index of a HI structure. Therefore, all the hasht- ables and the inverted indexes of HI can again be stored using a similar directory-file structure as $\mathbf{I}_{kp}$, and all the points in the dataset can be indexed into a B+-Tree [36] using their ids and stored on the disk. In this way, subset search can retrieve the points from the disk using B+-Tree for exploring the final set of results.

Acknowledgments

This research work is supported in part by the US National Science Foundation (NSF) under grant IIS- 1219254.

References

[1] W. Li and C. X. Chen, "Efficient data modeling and querying system for multi-dimensional spatial data," in *Proc. 16th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2008, pp. 58:1–58:4.

[2] D. Zhang, B. C. Ooi, and A. K. H. Tung, "Locating mapped resour- ces in web 2.0," in *Proc. IEEE 26th Int. Conf. Data Eng.*, 2010, pp. 521–532.

[3] V. Singh, S. Venkatesha, and A. K. Singh, "Geo-clustering of images with missing geotags," in *Proc. IEEE Int. Conf. Granular Comput.*, 2010, pp. 420–425.

[4] V. Singh, A. Bhattacharya, and A. K. Singh, "Querying spatial patterns," in *Proc. 13th Int. Conf. Extending Database Technol.: Adv. Database Technol.*, 2010, pp. 418–429.

[5] J. Bourgain, "On lipschitz embedding of finite metric spaces in hil- bert space," *Israel J. Math.*, vol. 52, pp. 46–52, 1985.

[6] H. He and A. K. Singh, "GraphRank: Statistical modeling and mining of significant subgraphs in the feature space," in *Proc. 6th Int. Conf. Data Mining*, 2006, pp. 885–890.

[7] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi, "Collective spatial keyword querying," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2011, pp. 373–384.

[8] C. Long, R. C.-W. Wong, K. Wang, and A. W.-C. Fu, "Collective spatial keyword queries: A distance owner-driven approach,' *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 689–700.

[9] D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung, and M. Kitsure- gawa, "Keyword search in spatial databases: Towards searching by document," in *Proc. IEEE 25th Int. Conf. Data Eng.*, 2009, pp. 688–699.

[10] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality- sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geometry*, 2004, pp. 253–262.

[11] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, "Hybrid index structures for location-based web search," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, 2005, pp.155–162.

[12] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, "Processing

spatial- keyword (SK) queries in geographic information retrieval (GIR) systems," in *Proc. 19th Int. Conf. Sci. Statistical Database Manage.*, 2007, p. 16.

[13] S. Vaid, C. B. Jones, H. Joho, and M. Sanderson, "Spatio- textual indexing for geographical search on the web," in *Proc. 9th Int. Conf. Adv. Spatial Temporal Databases*, 2005, pp.218–235.

[14] A. Khodaei, C. Shahabi, and C. Li, "Hybrid indexing and seamless ranking of spatial and textual features of web documents," in *Proc. 21st Int. Conf. Database Expert Syst. Appl.*, 2010, pp.450–466.

[15] A. Guttman, "R-trees: A dynamic index structure for spatial searching," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 1984, pp. 47–57.

[16] I. De Felipe, V. Hristidis, and N. Rishe, "Keyword search on spatial databases," in *Proc. IEEE 24th Int. Conf. Data Eng.*, 2008, pp. 656–665.

[17] G. Cong, C. S. Jensen, and D. Wu, "Efficient retrieval of the top-k most relevant spatial web objects," *Proc. VLDB Endowment*, vol. 2, pp. 337–348, 2009.

[18] B. Martins, M. J. Silva, and L. Andrade, "Indexing and ranking in Geo-IR systems," in *Proc. Workshop Geographic Inf.*, 2005, pp. 31–34.

[19] Z. Li, H. Xu, Y. Lu, and A. Qian, "Aggregate nearest keyword search in spatial databases," in *Proc. 12th Int. Asia- Pacific Web Conf.*, 2010, pp. 15–21.

[20] M. L. Yiu, X. Dai, N. Mamoulis, and M. Vaitis, "Top-k spatial pref- erence queries," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, 2007, pp. 1076–1085.

[21] T. Xia, D. Zhang, E. Kanoulas, and Y. Du, "On computing top-*t* most influential spatial sites," in *Proc. 31st Int. Conf. Very Large Databases*, 2005, pp. 946–957.

[22] Y. Du, D. Zhang, and T. Xia, "The optimal-location query," in *Proc. 9th Int. Conf. Adv. Spatial Temporal Databases*, 2005, pp. 163– 180.

[23] D. Zhang, Y. Du, T. Xia, and Y. Tao, "Progressive computation of the min-dist optimal-location query," in *Proc. 32nd Int. Conf. Very Large Databases*, 2006, pp. 643–654.

[24] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, "The R*- tree: An efficient and robust access method for points and rectangles," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 1990, pp. 322–331.

[25] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Int. Conf. Very Large Databases*, 1994, pp. 487–499.

[26] P. Ciaccia, M. Patella, and P. Zezula, "M-tree: An efficient access method for similarity search in metric spaces," in *Proc. 23rd Int. Conf. Very Large Databases*, 1997, pp.426–435.

[27] R. Weber, H.-J. Schek, and S. Blott, "A quantitative analysis and perfomance study for similarity-search methods in high-dimensional spaces," in *Proc. 24th Int. Conf. Very Large Databases*, 1998, pp. 194–205.

[28] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert Space," *Contemporary Math.*, vol. 26, pp. 189– 206, 1984.

[29] J. M. Kleinberg, "Two algorithms for nearest-neighbor search in high dimensions," in *Proc. 29th Annu. ACM Symp. Theory Comput.*, 1997, pp. 599–608.

[30] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. 25th Int. Conf. Very Large Data- bases*, 1999, pp. 518–529.

[31] V. Singh and A. K. Singh, "SIMP: Accurate and efficient near neighbor search in high dimensional spaces," in *Proc. 15th Int. Conf. Extending Database Technol.*, 2012, pp. 492–503.

[32] Y. Tao, K. Yi, C. Sheng, and P. Kalnis, "Quality and efficiency in high dimensional nearest neighbor search," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2009, pp.563— 576.

[33] H.-H. Park, G.-H. Cha, and C.-W. Chung,"Multi-way spatial joins using r-trees: Methodology and performance evaluation," in *Proc. 6th Int. Symp. Adv. Spatial Databases*, 1999, pp.229–250.

[34] D. Papadias, N. Mamoulis, and Y. Theodoridis, "Processing and optimization of multiway spatial joins using r-trees," in *Proc. 18th ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst.*, 1999, pp. 44–55.

[35] T. Ibaraki and T. Kameda, "On the optimal nesting order for computing N-relational joins," *ACM Trans. Database Syst.*, vol. 9, pp. 482–502, 1984.

[36] D. Comer, "The ubiquitous b-tree," *ACM Comput. Surveys*, vol.11, no. 2, pp. 121–137, 1979.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

An Efficient Method for High Quality and Cohesive Topical Phrase Mining

Anil
M.Tech Scholar,
CSE Department,
Malla Reddy College of Engineering
anil.suryapet@gmail.com

ABSTRACT

A phrase is a natural, meaningful, and essential semantic unit. In topic modeling, visualizing phrases for individual topics isan effective way to explore and understand unstructured text corpora. Usually, the process of topical phrase mining is twofold: phrase mining and topic modeling. For phrase mining, existing approaches often suffer from order sensitive and inappropriate segmentation problems, which make them often extract inferior quality phrases. For topic modeling, traditional topic models do not fully consider the constraints induced by phrases, which may weaken the cohesion. Moreover, existing approaches often suffer from losing domain terminologies since they neglect the impact of domain-level topical distribution. In this paper, we propose an efficient method for high quality and cohesive topical phrase mining. A high quality phrase should satisfy frequency, phraseness, completeness, and appropriateness criteria. In our framework, we integrate quality guaranteed phrase mining method, a novel topic model incorporating the constraint of phrases, and a novel document clustering method into an iterative framework to improve both phrase quality and topical cohesion. We also describe efficient algorithmic designs to execute these methods efficiently

INTRODUCTION

TOPICAL phrase mining refers to automatically extracting phrases which grouped by individual themes from given text corpora. It is of high value to enhance the power and efficiency to facilitate human to explore and understand a large amount of unstructured text data. One example is that if researchers could find phrases among a research field appearing with high frequencies in related proceedings in

different years, they will be able to have an insight into the academic trend of that research field. Topical phrase mining is not only an important step in established fields of

information retrieval and text analytics, but also is critical in various tasks in emerging applications, including topic detection and tracking [1], social event discovery [2], news recommendation system, and document summarization [3]. Usually, the process of topical phrase mining is twofold: phrase mining and topic modeling. These two stages not only directly affect the quality of discovered phrases and the cohesion of topics, but also, they may interact and indirectly impact each other's outcomes, e.g., low quality phrases (incomplete or meaningless) may cause misleading topical assignment in topic modeling. However, from phrase quality and topical cohesion perspectives, the outcomes of existing approaches remain to be improved. From phrase quality perspective, existing phrase mining methods [4–11] often produce low quality phrases. A high quality phrase should satisfy frequency, phrasegrness, completeness, and appropriateness criteria. Phrase mining is originated from the natural language processing (NLP) community, which utilizes predefined linguistic rules that rely on part-of-speech (POS) tagging or parsing trees [4, 5] to generate phrases. Such NLP based methods are commonly language-dependent and need texts to comply with grammar-rules, so it is not easy for them to be migrated to other languages and not suitable for analyzing some newly emerging and grammar-free text data, such as twitters, academic papers and query logs. In the hope to overcome the disadvantages of NLP based methods, there are many data-driven approaches that have been proposed in this area. Data-driven methods primarily view phrase mining as a frequent pattern mining problem [6, 7]. A phrase is extracted if it is constituted by the longest word sequence whose frequency is larger than a given threshold. Inevitably, extracting word sequence according to frequency is prone to produce many false phrases. Recently, researchers have sought for a kind of general, yet powerful phrase mining method. A variety of statistic-based methods [8–10] have been proposed to improve phrases quality by ranking

ISBN 978-93-88808-61-3

candidate phrases. A more recent work [11] considers integrating phrasal segmentation with phrase quality estimation to estimate rectified phrase frequency to further improve phrase quality.

However, due to suffering from order sensitive and inappropriate segmentation, the outcome of existing methods is still inadequate. Below we use Table 1 to show the deficiencies of the existing methods by using significance scores Sig score extracted from a corpus, 5Conf. 1 We compared two phrases using different processing orders based on 5Conf. Data in Table 1 is derived from the result of an existing method [9] which heuristically merges words under t-test score (i.e., a statistical hypothesis test to measure whether its actual occurrence significantly different from expected occurrence). The expected occurrence of phrase Pr = w1 _ w2 is calculated by f(w1)_f(w2) N , where f(w1) and f(w2) are word frequencies of w1 and w2 in the corpus, respectively, and N is the total number of words in the corpus. The method [9] allows users to specify a threshold of a significance score Sig score(Pr) of a phrase Pr, which is the statistical significance of taking a group of words as a phrase. It is measured by comparing the actual frequency with the expected occurrence. A larger value of Sig score(Pr) indicates the word sequence Pr has higher possibility to be a whole unit (phrase) than other sequences, and vice versa.

(1) Order sensitive. Assume Gaussian Mixture Model is a high quality phrase since it is complete in semantic. By choosing the merge order1 2 :3 , as shown in Table 1, existing approaches heuristically merge Gaussian and Mixture firstly, since the order shows a higher t-test score 6391:62 to achieve a local optimum comparing with the score 23:96 by using the order 2 3 :1 . However, if the threshold Sig score = 16, the complete phrase Gaussian Mixture Model failed to be extracted by using the order 1 2 :3 since the final core 15:75 is less than the given threshold 16 (we use symbol _ to

ISBN 978-93-88808-61-3

denote the score of the whole phrase under the given merge order). Instead, the merge order 2 3 :1 could have this phrase extracted. For the second phrase Peer to Peer Data, by using the same corpus, we got the same conclusion. Consequently, the completeness of extracted phrases highly depends on the merging order of the merging heuristics. The incompleteness brought by traditional approaches will cause incomplete semantics and may produce very general phrases. For instance, phrase Mixture Model has many explanations, such as Gaussian Mixture Model, Finite Mixture Model, or Interactive Mixture Model, whereas by phrase Gaussian Mixture Model, one explicitly refers to the very probabilistic model.

(2) Inappropriate segmentation. For the word sequence Gaussian Mixture Model Selection, it contains two quality phrases Gaussian Mixture Model and Model Selection since they both have high statistic scores. However, these two quality phrases are overlapping in the sequence. In the scenario of text chunking, the word

model can only belong to one of these two phrases, i.e., s1 = Gaussian Mixture Model | Selection or s2 = Gaussian Mixture | Model Selection. Existing approaches which only consider intra-coocurrence (e.g., phrase frequency and phrase length) prefer to choose sequence s2 , since both Gaussion Mixture and Model Selection have high frequencies. However, Gaussian Mixture Model should be the right choice for it is a whole function unit as an adjective, while Gaussian Mixture is obviously an incomplete phase.

From topical cohesion perspective, traditional topic models, such as LDA, assume words are generated independently from each other, i.e. “bag-of-words” assumption. Under this assumption, a phrase is regarded as an independent “word”, which may lead to the loss of its specific meaning, and as a result, the impact of phrases is ignored. To address the topic assignment problem associated with phrase, some existing methods

ISBN 978-93-88808-61-3

such as PhraseLDA [9] uses an undirected clique to model the stronger correlation of words in the same phrase on top of the "bag-of-phrases" assumption. To be specific, words in the same phrase form a clique, and PhraseLDA imposes the same latent topic on the words in the same clique. However, it is not enough to consider only the correlation of a phrase and its words. A phrase as a whole may carry lexical meaning that is beyond the sum of its individual words. For example, the phrase max pooling has a meaning beyong the word "max" or "pooling". Thus, it would be inappropriate to enforce words in the same phrase to inherit the same topic like PhraseLDA does, since long noun phrases sometimes do have components indicative of different topics [12].

Moreover, existing approaches neglect a fact that some phrases are only valid in certain domains. Usually, the texts within a corpus often come from more than one domain, and each domain may contain its own terminologies. These domain-specific terminologies may only appear frequently within certain domains but not in others, making them less possible to be extracted in the entire corpus where their occurrence frequency is diluted by the other domains, asTable 2 demonstrates.

In Table 2, the phrases support vector machine, eigen vector, bit vector, and social networks are estimated to belong to machine learning (ML) , math (MA), database (DB), and data mining (DM) domains, respectively. Even though some phrases (e.g., support vector machine and social networks) can achieve a high enough significance in the entire corpus, while others such as bit vector and eigen vector cannot. Consequently, it is hard for them to be mined as phrases in the entire corpus, albeit actually they both are common terminologies in their own domains.

Besides effectiveness, efficiency is also very important to topical phrase mining, especially for the applications that need timely analysis, such as topic-tracking [1], social event

ISBN 978-93-88808-61-3

discovery [2], and news recommendation system. Take Twitter as an example, the volume of tweets grew at increasingly high rates from its launch in 2006 to 2010, approaching around 1; 000% gain in yearly volume2. Currently, over 350; 000 tweets are generated on Twitter per minute. Unfortunately, most existing approaches [11–14] often suffer from low efficiency as they cannot support such high throughput tasks.

In order to effectively and efficiently mine topical phrases and improve phrase quality and topical cohesion, we propose a Cohesive and Quality Topical Phrase Mining (CQMine) framework, which automatically clusters documents with a more sensible topic model, and improves the quality of phrases by adopting more accurate and rigorous mining approaches. Moreover, our quality phrase mining approach can be solely used to mine phrases. The main contributions of this paper are as follows:

We propose effective and efficient quality phrase mining approaches. By eliminating order sensitive and avoiding inappropriate segmentation, our approaches could guarantee the quality of extracted phrases. Moreover, we also design effective algorithms to accelerate the processing.

We propose a novel topic model to address topic assignment problem associated with idiomatic phrases to improve the cohesion of topical phrases. Considering the fact that some phrases are only valid in certain domains, we propose an iterative framework to facilitate more accurate domain terminologies finding. _ Experimental evaluation and case study demonstrate that our method is of high interpretability and efficiency compared with the state-of-the-art methods.

Existing System

Topical phrase mining is not only an important step in established fields of information retrieval and text analytics, but also is critical in various tasks in emerging applications,

ISBN 978-93-88808-61-3

including topic detection and tracking , social event discovery , news recommendation system, and document summarization .the process of topical phrase mining is twofold: phrase mining and topic modeling. These two stages notonly directly affect the quality of discovered phrases and the cohesion of topics, but also, they may interact andindirectly impact each other's outcomes, e.g., low quality phrases (incomplete or meaningless) may cause misleading topical assignment in topic modeling. However, from phrase quality and topical cohesion perspectives, the outcomes of existing approaches remain to be improved.

NLP based methods are commonly language-dependent and need texts to comply with grammar-rules, so it is not easy for them to be migrated to other languages and not suitable for analyzing some newly emerging and grammar-free text data, such as twitters, academic papers and query logs. In the hope to overcome the disadvantages of NLP based methods, there are many data-driven approaches that have been proposed in this area. A variety of statistic-based methods have been proposed to improve phrases quality by ranking candidate phrases.

Proposed System

Considering the fact that some phrases are only valid in certain domains, we propose an iterative framework to facilitate more accurate domain terminologies finding. Experimental evaluation and case study demonstrate that our method is of high interpretability and efficiency compared with the state-of-the-art methods.

Future Work

Different with the existing model which only considers intra-co occurrence of phrases and regards the generation of segmentations as an independent process. Our methods comprehensively consider both the intra-co occurrence of phrases and the isolation of partition position. From a technical perspective, the isolation of "current" split position depends on the "future" generated split position. Thus, we need to check every possible new split positions to determine the isolation of current split position, which makes the computation of optimal segmentations very time consuming. To address this issue, we adopt a dynamic programming strategy, which is based on an observation that if bi+1 and the previous partition position bi is the optimal position.

News Publisher

News publisher provides the news articles on daily basis, breaking news; live news etc. news data are stored in

ISBN 978-93-88808-61-3

database. Offering the services to the end users. News Recommendation system publish the news articles based on categories. News Publisher search the news topics randomly whether the articles are displaying related to category. Users Registered in news portal to view the news articles, once read the article can also to comment the article and shared to others

Effectiveness Analysis of quality phrase

Examined the effectiveness of our quality phrase mining stage by measuring the phrase quality in two metrics: (1) Wiki-phrases benchmark and (2) Expert Evaluation. **Wiki-Phrases:** Wiki-phrases is a collection of popular mentions of entities by crawling intra-Wiki citations within Wiki content. Wiki phrases benchmark provides a good coverage of commonly used phrases which could avoid the variance caused by different human raters. In this evaluation,we regarded Wiki phrases as ground truth phrases. That is to belongs to/not belongs to Wiki phrases. To compute precision, only the Wiki phrases are considered to be positive. For recall, we firstly mergedall the phrases returned by all methods including ours, and then we obtained the intersection between the Wiki phrases and the merged phrases as the evaluation set.

Quality Phrase Mining

In the CQMine framework the quality phrase mining stage contains three steps:
Firstly, a PhraseTrie is built to count all possible phrases' frequencies. Then, a complete phrase mining algorithm is applied to mine complete phrases, which will be under the guidance of a statistics-based measurement to satisfy phraseness criterion. During phrase mining, the mined phrases are stored inPhraseTrie to avoid recomputing duplicate phrases. Finally, to guarantee the appropriateness requirement, for each document, CQMine needs to check if it contains overlapping phrases, if so, we will partition them into non-overlapping phrases by utilizing an effective and efficient overlaping phrases segmentation algorithm. After quality phrase mining, a document is transformed from a multiset of words (bag-of-words) into a multiset of phrases (bag-of-phrases) which will be taken as the input of topic modeling.

Topical phrase mining

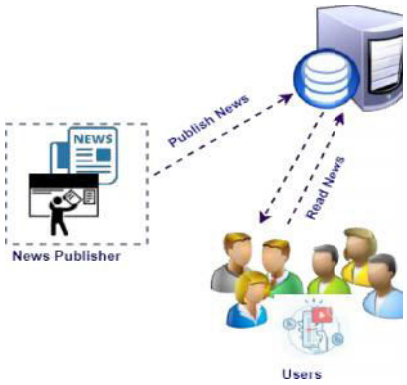
Significant progresses have been made on the topical phrase mining and they can be broadly classified into three types:

ISBN 978-93-88808-61-3

(1) Joint learning phrases and their topic assignment,
(2) Mining phrases posterior to topic inferring,
(3) Mining phrases prior to topic inferring.

Word sequence segmentation (or phrasal segmentation) is another strategy for phrase mining. Formally, phrasal segmentation aims at partitioning a word sequence into a set of disjoint subsequences, each indicating a phrase. It only considers intra co occurrence of phrases such as phrase length and words, while ignores the inter-isolation between phrases. The second strategy utilizes a post-processing step to generate phrases after inferred by the LDA model. Recursively merges consecutive words with the same latent topic by a distribution-free permutation test on arbitrary length back-off model until all significant Consecutive words have been merged. it performs phrase mining and topic inferring simultaneously by incorporating successive word sequence assumption into the generative model. Wallach proposed a bigram topic model based on a hierarchical Dirichlet allocation model. Bigram model is a probabilistic generative model that conditions on the previous word and topic when drawing the next word.

Architecture



Algorithm

The completeness of extracted phrases highly depends on the merge order. In order to obtain the complete phrases, we need to enumerate every possible merge order. Obviously, a straight-forward algorithm of finding the complete phrases in document d is: enumerating all the subsequences of this document first, then verify whether each one is a complete phrase.The algorithm QBA (q-Chunk Based Approach) firstly generates boundaries It then computes the local solution of each chunk using DPBA denote the left boundary of current chunk. For each boundary algorithm QBA checks whether satisfies merge condition.

The main processing steps of QBA are as follows:
(1) Partitioning the sequence into a series of q-length chunks;
(2) Performing top-down search on each chunk to get local solutions

(3)Checking whether two adjacent chunks need to be merged.

If they do not need to be merged, it means no phrase could cross the boundary between the two chunks. Otherwise the two chunks are merged into a new chunk and QBA will find new solutions on the new chunks.

CONCLUSIONS

We presented an efficient method for cohesion and quality topical phrase mining. In phrase mining stage, we focus on quality phrase mining problem, and propose two efficient quality phrase mining algorithms. In practice, the time cost of our best exact algorithm is competitive to greedy algorithm. In topic modeling stage, we propose a novel topic model to incorporate the constraint that is induced by phrases; moreover, it can well address the collocation phrase issue. Finally, considering the fact that some phrases are only valid in certain domains, we cluster documents under the condition that they share similar topic distribution and iteratively perform cluster updating and topical inferring to further improve the cohesion of topical phrases. The empirical verification demonstrated our framework has high interpretability and efficiency.

REFERENCES

1. J. Leskovec, L. Backstrom, J. Kleinberg, "Meme-tracking and the dynamics of the news cycle", *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 497-506, 2009.

2. M. Li, J. Wang, W. Tong et al., "EKNOT: Event knowledge from news and opinions in twitter", *Proc. 30th AAAI Conf. Artif. Intell.*, pp. 4367-4368, 2016.

3. Z. He, C. Chen, J. Bu et al., "Document summarization based on data reconstruction", *Proc. AAAI Conf. Artif. Intell.*, pp. 620-626, 2012.

4. S. P. Abney, "Parsing by chunks" in Principle-Based Parsing, The Netherlands:Kluwer Academic Publishers, pp. 257-278, 1991.

5. H. Clahsen, C. Felser, "Grammatical processing in language learners", *Applied Psycholinguistics*, vol. 27, no. 27, pp. 3-41, 2006.

6. M. Danilevsky, C. Wang, N. Desai et al., "Automatic construction and ranking of topical keyphrases on collections of short documents", *Proc. Int. Conf. Data Mining*, pp. 398-406, 2014.

7. A. Simitsis, A. Baid, Y. Sismanis et al., "Multidimensional content exploration", *Proc. VLDB Endowment*, vol. 1, no. 1, pp. 660-671, 2008.

8. A. Parameswaran, H. Garcia-Molina, A. Rajaraman, "Towards the web of concepts: Extracting concepts from large datasets", *Proc. VLDB Endowment*, vol. 3, no. 1–2, pp. 566-577, 2010.

9. A. El-Kishky, Y. Song, C. Wang et al., "Scalable topical phrase mining from text corpora", *VLDB Endowment*, vol. 8, no. 3, pp. 305-316, 2014.

10. P. Deane, "A nonparametric method for extraction of candidate phrasal terms", *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics*, pp. 605-613, 2005.

11. J. Liu, J. Shang, C. Wang et al., "Mining quality phrases from massive text corpora", *Proc. ACM SIGMOD Int. Conf. Manage. Data*, pp. 1729-1744, 2015.

12. X. Wang, A. McCallum, X. Wei, "Topical N-grams: Phrase and topic discovery with an application to information retrieval", *Proc. 7th IEEE Int. Conf. Data Mining*, pp. 697-702, 2007.

13. C. Wang, M. Danilevsky, N. Desai et al., "A phrase mining framework for recursive construction of a topical hierarchy", *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 437-445, 2013.

14. R. V. Lindsey, W. P. Headden, M. J. Stipicevic, "A phrase discovering topic model using hierarchical pitman-yor processes", *Proc. Joint Conf. Empirical Methods Natural Language Processing Comput. Natural Language Learn.*, pp. 214-222, 2012.

15. E. Pitler, S. Bergsma, D. Lin et al., "Using web-scale N-grams to improve base NP parsing performance", *Proc. 23rd Int. Conf. Comput. Linguistics*, pp. 886-894, 2010.

16. M. F. Porter, Snowball: A Language for Stemming Algorithms, Palo Alto, CA, USA:Open Source Initiative Osi, 2001.

17. M. F. Porter, "An algorithm for suffix stripping", *Program*, vol. 14, no. 3, pp. 130-137, 1980.

ISBN 978-93-88808-61-3

18. K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *Philosophical Magazine*, vol. 50, no. 302, pp. 157-175, 1900.

19. T. L. Griffiths, M. Steyvers, "Finding scientific topics", *Proc. Na. Academy Sci. United States America*, vol. 101, no. 1, pp. 5228-5235, 2004.

20. S. Geman, D. Geman, "Stochastic relaxation gibbs distributions and the Bayesian restoration of images", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721-741, Nov. 1984.

21. B. Li, B. Wang, R. Zhou et al., "CITPM: A cluster-based iterative topical phrase mining framework", *Proc. Int. Conf. Database Syst. Advanced Appl.*, pp. 197-213, 2016.

22. S. Kullback, R. A. Leibler, "On information and sufficiency", *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79-86, 1951.

23. A. Rodriguez, A. Laio, "Clustering by fast search and find of density peaks", *Sci.*, vol. 344, no. 6191, pp. 1492-1496, 2014.

24. K. Frantzi, S. Ananiadou, H. Mima, "Automatic recognition of multi-word terms: The C-value/NC-value method", *Int. J. Digital Libraries*, vol. 3, no. 2, pp. 115-130, 2000.

25. I. H. Witten, G. W. Paynter, E. Frank et al., "KEA: Practical automatic keyphrase extraction", *Proc. ACM Conf. Digital Libraries*, pp. 254-255, 1999.

Proceedings on International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Nearest Keyword Set Search Queries on Multi-Dimensional Datasets

Mr.Rajesh,
M.Tech.
Scholar,
Department of CSE,
Malla Reddy College of Engineering
rajcherukumalli@gmail.com

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 queryin 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

ISBN 978-93-88808-61-3

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 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.

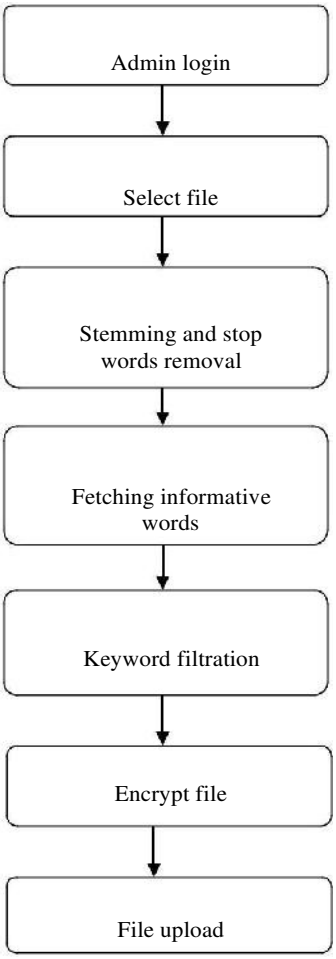


Figure 1 Work flow of admin

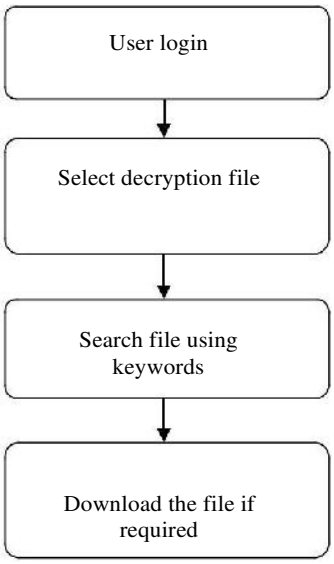


Figure 2 Work flow of user

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 978-93-88808-61-3

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.

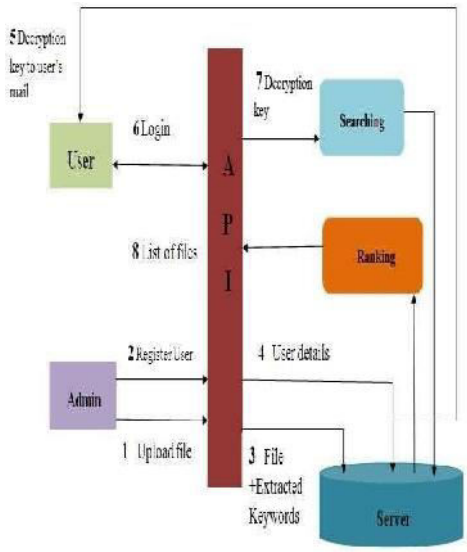
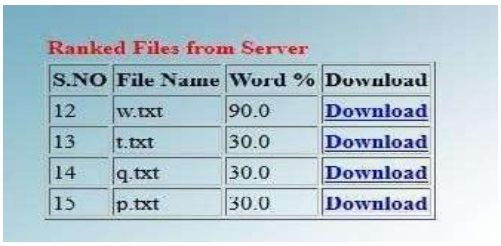


Figure 3: System architecture

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.

ISBN 978-93-88808-61-3

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 International Conference on Emerging Technologies in Computer Science(ICETCS-2019)
ISBN 978-93-88808-61-3

Analyzing And Detecting Money-Laundering Accounts in online social networks

Thati
Sindura
M.Tech
Scholar
Malla Reddy College of Engineering Hyderabad

Abstract

Virtual currency in OSNs plays an increasing- ly important role in supporting various financial activities such as currency exchange, online shop- ping, and paid games. Users usually purchase virtual currency using real currency. This fact motivates attackers to instrument an army of accounts to collect virtual currency unethically or illegally with no or very low cost and then launder the collected virtual money for massive profit. Such attacks not only introduce significant financial loss of victim users, but also harm the viability of the ecosystem. It is therefore of central importance to detect malicious OSN accounts that engage in laundering virtual currency. To this end, we extensively study the behavior of both malicious and benign accounts based on operation data collected from Tencent QQ, one of the largest OSNs in the world. Then, we devise multi-faceted features that characterize accounts from three aspects: account viability, transaction sequences, and spatial correlation among accounts. Finally, we propose a detection method by integrating these features using a statistical classifier, which can achieve a high detection rate of 94.2 percent at a very low false positive rate of 0.97 percent.

Introduction

Online social networks (OSNs) have started to leverage virtual currency as an effective means to glue financial activities across various platforms such as online shopping, paid online games, and paid online reading. Examples of virtual currency in such OSNs include but are not limited to Ten- cent Q Coin, Facebook Credits[1], and Amazon Coin. Usually, users purchase virtual money using real currency at a regulated rate; one user can also transfer it to another user via various meth- ods such as recharging their account and send- ing gifts [1]. These facts enable attackers to gain potentially massive profits through the following steps. First, an attacker can collect virtual currency with zero or low cost. For example, they can com- promise and subsequently control a legitimate account or register a huge number of accounts to win gifts (in the form of virtual currency) in online promotion activities. Next, they can instru- ment accounts under their control to transfer vir- tual currency to other accounts in return for real currency, with rates that are usually much lower compared to the regulated rate. Attackers usually post advertisements in popular e-commerce web- sites [2] to attract potential buyers. We call OSN accounts that are used by attackers for the collec- tion and transfer of virtual currency *MONEY-LAUN- dering ACCOUNTS*. Money- laundering accounts have caused a tremendous financial loss for compro- mised accounts, fundamentally undermined the effectiveness of online promotion activities, and possibly introduced potential conflicts against cur- rency regulations.

Detecting money-laundering accounts in OSNs therefore becomes of essential importance, which, however, is faced with new, significant challenges. First, committing money- laundering activities does not require the use of traditional malicious content such as spam, malicious URLs, or malicious executables. Although spamming might be used by attackers for advertising, neither methods nor the accounts used for spamming are necessarily

associated with the money-launder- ing accounts. Second, money-laundering activi- ties do not rely on social behavior and structures (e.g., "following" or "friend" relationship in popu- lar social networks) to operate. These challenges make existing methods immediately ineffective, since they focus on detecting OSN-based spam- ming, phishing, and scamming attacks, whose proper operation necessitates malicious content [3, 4], social structures [5], or social behaviors [6]. Detecting money laundering activities in tra- ditional financial transactions has attracted signif- icant research efforts [7]. For example, Dreewski *et al.* [8] designed a system to detect money laun- dering activities from billings and bank account transactions. Paula *et al.* [9] used the AutoEncod- er to classify exporters and detect money launder- ing activities in exports of goods and products in Brazil. Colladon *et al.*
[10] presented predictive models to quantify risk factors of clients involved in the factoring business and proposed a visual analysis method to detect the potential clusters of criminals and prevent money laundering. Dif- ferent from traditional money laundering detec- tion problems in bank-related activities, account behaviors of laundering virtual currency in OSNs involve bank-related financial activities, online social networks, and virtual recharging and expen- diture activities.

The goal of our work is to design an effective method capable of detecting money- laundering accounts. As a means toward this end, we per- form an extensive study of behaviors of mon- ey-laundering accounts based on data collected from Tencent QQ, one of the largest OSNs in the world with a giant body of reportedly 861 million active users. We have devised multi-facet- ed features that characterize accounts from three aspects: account viability, transaction sequences, and spatial correlation among accounts. Experi- mental results have demonstrated that our meth- od can achieve a high detection rate of
94.2 percent with a very low false positive rate of 0.97
percent. To the best of our knowledge, this work represents the first effort to analyze and detect money-laundering accounts in OSNs integrating virtual currency at this large scale.

Data Set

We have collected labeled data from Tencent QQ, a leading online social network in China, which offers a variety of services such as instant messaging, voice chat, online games, and online shopping. These services are glued together using Q coin, the virtual currency distributed and man- aged by Tencent QQ. Tencent QQ has a giant body of 861 million active accounts with a report- ed peak of 266 million simultaneously online users. Also, Tencent QQ is one of the leading OSNs that are actively involved in virtual currency based services in the world.

Our data set is composed of 114,891 mali- cious accounts and 381,523 benign accounts that are active during the first week of August in 2015. In order to label accounts used for money laundering, we follow advertisements of cheap virtual currency in major e-commerce websites and actually purchased virtual currency from sell- ers, where QQ accounts used by these sellers are labeled as money- laundering accounts. Since an attacker usually controls a large number of mali- cious accounts for money laundering, we label accounts as malicious if they login from the same IP address used by a confirmed money-laundering account within one day.

Although this labeling process offers us the ground truth, using it as a detection method is practically challenging. First, it requires a consider- able amount of investment to engage money-laun- dering accounts in malicious activities. Second, the IP addresses used to label launder accounts usually will be invalid after a few days, because attackers change the

ISBN 978-93-88808-61-3

login IP addresses frequent- ly. Therefore, this data labeling process, if used as a detection method, cannot guide OSNs to mitigate their financial loss proactively.For each account, we collect the following activity records. It is worth noting that all these records can be col- lected from social networks that integrate virtual currency.

- Login activities, which include the account ID, the login date, the login IP address, and the account level.
- Expenditure activities, which include the expenditure account ID, the expenditure date, the expenditure amount, the pur- chased service, the payment method, and the account ID to receive the service.
- Recharging activities, which include the recharging account ID, the recharging date, the recharging amount, and the payment method.

Behavior analysis and feature extraction

Figure 1 shows a typical process of virtual currency laundering. The first step is to collect virtual currency with zero or extremely low cost. For example, attackers can hack users' accounts (and thus control their virtual currency), exploit the system vulnerabilities, or participate in online promotion activities to win virtual currency for free or at significantly discounted rates [2]. Next, attack- ers attract potential buyerswith considerable

centage of benign accounts (less than 20 percent) experience the same activity level (i.e., being activefor less than 10 percent of total days).

Next, we study the source of virtual currency for benign and laundering accounts. A benign user usually recharges their account via wire trans- fer (often in the form of mobile payment) and occasionally receives gifts (from friends). Comparatively, money-laundering accounts almost exclusively rely on online promotions to directly collect virtual currency or gifts transferred from other accounts. We therefore introduce the following feature to characterize the currency collection behavior.

FeatUre 3: Percentage of Recharge from Mobile Payment: This feature represents the per- centage of virtual currency recharged through mobile payments (i.e., purchasing virtual currency using mobile online banks).

Figure 2c presents the distribution for this feature, where approximately 24 percent of benign users recharge their accounts via mobile payment, while the vast majorit y of malicious accounts do not use this channel.

As an increasing number of financial functions areintegrated into social networks, usersconduct a variety of activities such as shopping and gift- ing. While benign users prefer to engage financial activities with higher diversity, money- laundering accounts only focus on activities relevant to laundering. Therefore, we introduce the following five features to characterize such a difference.

FEATUre 4: The NUMber of Self-EXPENDITUREs: This feature represents the total number of expenditures that an account has committed to itself using virtual currency.

FEATURE 5: The NUMBER of EXPENDITURES Not from VIRTUAL CURRENCY: This feature characteriz- es the number of expenditures an account has committed by other methods instead of virtual currency.

FeatUre 6: Percentage of ExpenditUre from Banks: A user can associate their bank account with the OSN account. This bank account can be directly used for shopping and gifting in addition to virtual currency in the OSN account. This feature is defined as the percentage of expenditure from associated bank accounts.

FeatUre 7: The NUmber of AccoUnts that ever Receive Gifts from this ACCOUNT:

ISBN 978-93-88808-61-3

Malicious accounts need to frequently transfer the virtual currency as a gift to the buyer accounts, while a benign user tends to expend the virtual currency themself, and occasionally gives the virtual cur- rency as a gift to their friends. Thus, malicious accounts will have a much larger value of this fea- ture than benign users.

Feature 8: Percentage of the Amount of Expenditures as Gifts: This feature represents the proportion of the amount of expenditures as gifts in all expenditures. After malicious accounts collect virtual currency from the online promo- tion activities and other vulnerabilities, they will transfer it to other accounts as gifts. We therefore introduce this feature to quantify the percentage of all giving out behavior.

Figures 2d–2h report the distributions for Features 4–8 respectively. Almost all the malicious accounts (more than 99 percent) neither committed for itself using virtual currency nor committed by other methods instead of virtual currency. Comparatively,
61 percent of benign accounts have committed for itself using virtual currency at least once, and 18 percent of benign accounts have committed by other methods instead of vir- tual currency at least once. The distributions for Features 6–8 are also distinguishable as shown in the figures. We omit the descriptions for brevity.

Sequential AI Features of Financial Activities

The sequences of financial activities are likely to differ between benign accounts and money- laundering accounts. In order to model the sequential behavior, we use the discrete-time Markov Chain model. Specifically, we record the sequence of three basic financial activities: virtual-currency recharge, self- expenditures, and expenditures as gifts. Each state in the Markov Chain corresponds to one activity and the transition between two states represents a pair of two consecutive finan- cial activities. Hence, the Markov Chain has three states and nine total transitions. Each transition is associated with the probability of this transition among all observed transitions. Figure 3a illustrates how Markov Chain models are derived from a sequence of financial activities. Specifically, nodes 1', 2', and 3' refer to the three states "virtual- currency exchange," "self-expenditure," and "expenditure as gifts"; P_{ij} denotes the transition probability from state i to state j.

Figure 3b presents the CDF of P_{11}, P_{31}, and P_{33} for malicious accounts (denoted as "MA") and benign accounts (denoted as "BA"), respectively. As shown in the empirical analysis, the values of P_{11} and P_{33} for malicious accounts are much larger than those for benign accounts, which indicates that malicious accounts are more inclined to exchange multiple times continuously (see P_{11}), and expend as gifts multiple times continuously (see P_{33}). The values of P_{31} of malicious accounts are much smaller than those of benign accounts, which implies that benign accounts are more active to recharge virtual currency after expending as gifts compared to malicious accounts. It is worth noting that we omit the other six transition probabilities in the figure for brevity.

Our empirical analysis demonstrates that the sequential behaviors indeed experience significant differences between malicious and benign accounts. Therefore, we define the following features.

Features 9–17: The transition probabilities P_{11}, P12, P13, P21, P22, P23, P31, P32, P33.

Features 18–47: The top 30 most effective subsequences mined from the sequence of financial activities for malicious accounts. To achieve an acceptable time complexity, the

ISBN 978-93-88808-61-3

PrefixSpan algorithm [11] is used to mine the frequent subse- quences of behavior sequences. Then, the effec- tiveness e of mined subsequence q is measured by Eq. 1. In the equation, f_q denotes the number of times that subsequence q occurs in all the is associated with the probability of this tran- sition among all observed transitions. Figure 3a illustrates how Markov Chain models are derived from a sequence of financial activities. Specifical- ly, nodes 1', 2', and 3' refer to the three states "virtual-currency exchange," "self-expenditure," and "expenditure as gifts"; P_{ij} denotes the transi- tion probability from state i to state j.

Figure 3b presents the CDF of P_{11}, P_{31}, and P_{33} for malicious accounts (denoted as "MA") and benign accounts (denoted as "BA"), respectively. As shown in the empirical analysis, the values of P_{11} and P_{33} for malicious accounts are much larger than those for benign accounts, which indicates that malicious accounts are more inclined to exchange multiple times continuously (see P_{11}), and expend as gifts multiple times continuously (see P_{33}). The values of P_{31} of malicious accounts are much smaller than those of benign accounts, which implies that benign accounts are more active to recharge virtual currency after expending as gifts compared to malicious accounts. It is worth noting that we omit the other six transition probabilities in the figure for brevity.

Our empirical analysis demonstrates that the sequential behaviors indeed experience significant differences between malicious and benign accounts. Therefore, we define the following features.

FEATURES 9–17: The transition probabilities P_{11}, P12, P13, P21, P22, P23, P31, P32, P33.

FeatUres 18–47: The top 30 most effective subsequences mined from the sequence of financial activities for malicious accounts. To achieve an acceptable time complexity, the PrefixSpan algorithm [11] is used to mine the frequent subsequences of behavior sequences. Then, the effectiveness e ofmined subsequence q ismeasured by Eq. 1. In the equation, f_q denotes the number of times that subsequence q occurs in all the spAtIAl FeAtures oF currency trAnsFer

Each transaction for currency transfer can be characterized as a tuple denoted as $<s, t>$, where s and t refer to the source and destination account, respectively. For a node s, we identify a set of nodes, to each of which s has transferred virtu- al currency. We denote this set of nodes as $D(s)$ for this node$s$.We then

one or a few accounts (e.g., as birthday gifts) and thus form a fully connected graph, whose edges, however, are likely to have small weights. Since an account may receive gifts from both benign accounts (e.g., friend accounts) and money-laundering accounts, edges that connect benign and money- laundering accounts will also exist. To summarize, the graph is mainly composed of three types of connected subgraphs: subgraphs entirely composed of fully- connected malicious accounts, subgraphs entirely composed of fully-connected benign accounts, and subgraphs composed of both malicious accounts and benign accounts. Figure 4a presents one example of the third type of connected subgraphs. Specifically, malicious accounts A-D and C-E transfer to the same destination accounts respectively, a destination account obtains the virtual currency from both malicious account E and benign account F, and benign accounts F-I transfer to the same desti- nation account. Then, the corresponding graph is a connected graph composed of both malicious accounts and benign accounts.

ISBN 978-93-88808-61-3

Through analyzing the behaviors of destina- tion accounts, we find that most of the destination accounts as buyers tend to purchase the virtual currency or goods from the launder accounts rather than receiving gifts from benign accounts, and other destination accounts behave in the oppo- site way. This finding is validated by analyzing the neighbors of malicious accounts and benign accounts in the graph. It is analyzed that 80.1 percent of the neighbors of malicious accounts are malicious and 84.3 percent of the neighbors of benign accounts are benign on average. Thus, the malicious accounts and benign accounts tend to connect with the same type of vertices, and form a community structure in which some dense- ly connected components are composed of the same type of vertices and the connections among the components are sparse. The interpretation of the forming of community structure in transferring related graph is shown in Fig. 4a, and a real illustration of the structure is shown in Fig. 4b, where the red vertex denotes a malicious account and the blue vertex denotes a benign account.

To design the spatial features, we process the behaviors of accounts in the following two steps.

Step 1: Form the graph based on the definition of $G(V, E)$.

Step 2: Detect the densely-connected sub- graphs (communities) of the connected subgraphs of G based on a widely-used community detec- tion method, Fast Unfolding [12]. The method is a heuristic method based on modularity optimization, and is capable of dealing with large weighted graphs due to its acceptable time complexity.

Following the above two steps, the graph G will be divided into many communities, each account will belong to a community, and each community will be composed of almost the same type (malicious or benign) of accounts. We pres- ent the features of each account (vertex) below.

FEATURES of General ATTRIBUTES of Vertex in Graph

FeatUre 48 — Degree: The number of con-

detectIon And evAluAtIon

We leverage machine learning techniques to inte- grate all these features toperform effective detec- tion. Specifically, feature values extracted from labeled malicious and benign users have been employed to train a statistical classifier. After an unknown user is represented by a vectorof fea-

ture values, the classifier can automatically eval-

In order to evaluate the effectiveness of the proposed detection method, we use a total number of 496,414 accounts, of which 114,891 are malicious and 381,523 are benign. Without the loss of generality, we use Support Vector Machine(SVM), Random Forest (RM), and Logistic Regression (LR) [14] as the statistical classifier, where the SVM classifier was trained with a Gaussian Kernel and the RF classifier was trained with 3000 trees. We use three metrics to quantify the effectiveness of our method: detection rate (same definition as the true positive rate), false positive rate (FPR), and the area under the ROC curve (AUC) [15]. Specifically, AUC is a widely-used measure of the quality of the statistical classifier. It is defined as the probability that a randomly chosen sample of malicious accounts will have a higher estimated probability of belonging to malicious accounts than that of benign accounts. Since AUC is cutoff-independent and the values of AUC range from 0.5 (no predictive ability) to 1.0 (perfect predictive ability), a higher AUC of a classifier indicates better prediction performance, irrespective of the cutoff selection.

We perform 10-fold cross-validation to evaluate the detection performance of each select- ed statistical classifier based on all features,

ISBN 978-93-88808-61-3

using metrics including DR, FPR, and AUC. The results are presented in Table 1. Both Support Vector Machine and Random Forest can achieve high detection rates, high AUC values, and very low false positive rates. These results demonstrate that the features we extract can effectively differentiate between malicious accounts and benign accounts.

We evaluate the effectiveness of our meth- od when using features from one aspect or two aspects. Table 1 presents the results when SVM is adopted as the statistical classifier. The experimen- tal results demonstrate that features from each aspect show great promise in effectively detecting malicious accounts; features of two aspects show better performance compared to features from one aspect; the integration of features from all three aspects show the best performance. This

TABLE 1. Performance analysis of the detection method.

Classifiers	Features	FPR	Detection rate	AUC
SVM	All features	0.97 %	94.2%	0.966
RF	All features	0.22	92.3	0.960
LR	All features	4.56	90.2	0.928
SVM	Vitality features	3.0	86.9	0.920
SVM	Sequential features	3.83 %	93.3%	0.947
SVM	Spatial features	2.4	91.6	0.946
SVM	Vitality + sequential features	1.47	92.9	0.957
SVM	Vitality + spatial features	1.64	93.7	0.961
SVM	Sequential + spatial features	1.38 %	94.0%	0.963

implies high robustness of the proposed method. Specifically, if features of one aspect are evaded by attackers, remaining features can still accom- plish high detection accuracy.

On the scalability of the proposed detection method, although some of the vitality features may not be suitable for all the social networks (e.g., Feature 3 — percentage of rechargefrom mobile payment, because of that not all the social networks support mobile payment), the sequential and spatial features can be extracted in almost all the social networks integrating vir- tual currency, and are effective enough to detect the malicious accounts according to the per- formance analysis shown in Table 1. Therefore, other social networks can also adopt and extend the proposed method to detect the money-laun- dering accounts.

We also analyze the contribution ofeach sin- gle feature using information gain, where a higher value of information gain indicates more signifi- cant contribution. The rank of each feature based on information gain is shown in Table 2, where the top 20 features consist of seven spatial fea- tures, eight sequential features, and five vitality features. This indicates that all three aspects are useful for detection.

Conclusions

This article presents the analysis and detection method of money-laundering accounts in OSNs. We analyzed and compared the behavior of both malicious and benign accounts from three perspectives: the account viability, the transaction sequences, and spatial correlation among accounts. We designed a collection of 54 features to systematically characterize the behavior of benign accounts

and malicious accounts. Experimental results based on labeled data collected from Tencent QQ, a global leading OSN, demonstrated that the proposed method achieved high detection rates and very lowfalse positive rates.

References

[1] Y. Wang and S. D. Mainwaring, "Human-Currency Interaction: Learning from Virtual Currency use in China," *Proc. SIGCHI Conf. HUMAN Factors in COMPUTING Systems*, ACM, 2008, pp. 25–28.

[2] Y. Zhou *et al*., "ProGuard: Detecting Malicious Accounts in Social-Network-Based Online Promotions," *IEEE Access*, vol. 5, 2017, pp. 1990–99.

[3] F. Wu *et al*., "Social Spammer and Spam Message Co-Detection in Micro blogging with Social Context Regularization," *Proc. 24th ACM Int'l. Conf. Information and Knowledge Man-agement*, ACM, 2015, pp. 1601–10.

[4] L. Wu *et al*., "Adaptive Spammer Detection with Sparse Group Modeling," *Proc. 11th Int'l. AAAI Conf. Web and Social Media*, AAAI, 2017, pp. 319–26.

[5] S. Fakhraei *et al*., "Collective Spammer Detection in Evolving Multi-Relational Social Networks," *Proc. 21st ACM SIGKDD Int'l. Conf. Knowledge Discovery and Data Mining*, ACM, 2015, pp. 1769–78.

[6] F. Hao *et al*., "Robust Spammer Detection in Microblogs: Leveraging User Carefulness," *ACM Trans. Intelligent Sys- tems and Technology*, vol. 8, no. 6, 2017, pp. 83:1–31.

[7] G. K. Palshikar, "Detecting Frauds and Money Laundering: A Tutorial," *Proc. Int'l. Conf. Big Data Analytics*, Springer, 2014, pp. 145–60.R. Dreewski, J. Sepielak and W. Filipkowski, "The Application of Social Network Analysis Algorithms in a System Support- ing Money Laundering Detection," *Information Sciences*, vol. 295, 2015, pp. 18–32.

[8] E. L. Paula *et al*., "Deep Learning Anomaly Detection as Sup- port Fraud Investigation in Brazilian Exports and Anti-Money Laundering," *2016 15th IEEE Int'l. Conf. Machine Learning and Applications (ICMLA*), Anaheim, CA, 2016, pp. 954–60.

[9] A. F. Colladon and E. Remondi, "Using Social Network Analysis to Prevent Money Laundering," *Expert Systems with Applications*, vol. 67, 2017, pp. 49–58.

[10] J. Pei *et al*., "Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach," *IEEE Trans. Knowledge and Data Engineering*, vol. 16, no. 11, 2004, pp. 1424–40.

[11] M. E. J. Newman, "Communities, Modules and Large-Scale Structure in Networks," *NATURE Physics*, vol. 8, no. 1, 2012, pp. 25–31.

[12] R. Li *et al.*, "Finding Influential Communities in Massive Net- works," *The VLDB JOURNAL*, 2017.

[13] S. Rogers, and M. Girolami, *A First CoUrse in Machine Learning*, CRC Press, 2016.

[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and TECHNIQUES*, Elsevier, 2011.

ABOUT MRGI

Malla Reddy group of Institutions is one of the biggest conglomerates of hi-tech professional educational institutions in the state of Telangana, established in 2001 sprawling over 200 acres of land. The group is dedicated to impart quality professional education like pharmacy, Engineering & Technology, MCA, MBA courses. Our sole objective is to turn out high caliber professionals from those students who join us.

ABOUT MRCE

Malla Reddy group of Engineering (Formerly CM Engineering College) has been established under the aegis of the Malla Reddy Group of Institutions in the year 2005, a majestic empire, founded by chairman Sri Ch.Malla Reddy Garu. He has been in the field of education for the last 23 years with the intention of spearheading quality education among children from the school level itself. Malla Reddy College of Engineering has been laid upon a very strong foundation and has ever since been excelling in every aspect. The bricks of this able institute are certainly the adept management, the experienced faculty, the selfless non-teaching staff and of course the students.

ABOUT ICTIMES

ICTIMES started long back with its banner to promote the vision of future technologies that change the trends of life on this planet earth. Under this banner, the Department of Computer Science and Engineering at MRCE organizes the ICETCS – International Conference on Emerging Technologies in Computer Science to provide a scholarly platform to ignite the spirit of Research and bring out the latent potential in teaching fraternity and student community. ICETCS accommodates major areas like, Big Data , Data mining, Information Retrival, Neural Networks, Data Security, etc.,

ABOUT ICETCS

International Conference on Emerging Technologies in Computer Science ((ICETCS -2019) will bring together innovative academicians, researchers and industrial experts in the field of Computer Science to a common forum. The idea of the conference is for the scientists, scholars, engineers and students from the Universities across the world and the industry as well, to present ongoing research activities, and hence to foster research relations between the Universities and the industry with the rapid development of trends and studies in the fields concerned. ICETCS-2019 will provide a heartwarming platform to researchers, scholars, faculty and students to exchange their novel ideas face to face togethers.

Estd :2005

MALLA REDDY COLLEGE OF ENGINEERING

Approved by AICTE - New Delhi, Affiliated to JNTU - Hyderabad,Accredited by NBA & Accredited by NAAC.
ISO 9001:2015 Certified Institution, Recognition of College under Section 2(f) & 12 (B) of the UGC Act,1956.

Address: Maisammaguda, Dhulapally,(Post Via Kompally),Secunderabad - 500 100.

Ph: 040-64632248, 9348161222 ,9346162620. Email: principal@mrce.in

Website : www.mrce.in