International Journal of Advanced Information in Engineering Technology (IJAIET) ISSN: 2454-6933
Vol.3, No.1, January 2016

A Data Mining Approach for Secure Chat In IM

Prajeesh C B.
Department of Computer Science And Engineering
MEA Engineering College Perinthalmanna
Kerala, India.

Afsar P.
Department of Computer Science And Engineering
MEA Engineering College Perinthalmanna
Kerala, India

***Abstract*-Individuals and organizations alike depend on cyberspace. From the casual consumer using social media, to online merchants growing their business, to physicians supporting their patients every sector of our society is increasingly dependent upon technology and networked systems. Without sufficient awareness of the risks in cyberspace, however, Behavioural decisions and unseen threats can negatively impact the security of the global cyberspace infrastructure and can cause physical damage in the real world. On an individual level, what is at stake is the vulnerability of each individual user in cyberspace. An individual who is not aware of, and does not implement, basic cyber security practices faces greater personal risk on and online, such as identity theft, when engaging in daily online tasks. Through the Secure Chat System, it will counter these risks and help make the cyberspace, especially Instant Messenger (IM) more secure. Secure Chat can prepare the general public to identify and avoid risks in instant messenger.**

Keywords—Secure Chat, Instant Messenger, Ant phishing, deceptive phishing, cyber security, data mining, phishing.

I. INTRODUCTION

Our Nation's growing dependence on cyberspace is evident all around us. From smart phones and online banking to electronic health records, social networking, and automated manufacturing, our Nation increasingly relies on cyberspace. The scientists and innovators of tomorrow also rely on cyberspace to make the discoveries and inventions that improve our lives and drive our economy. The need for a safe and secure cyberspace has never been more important.

While there is no doubt that technology has changed the way we live, work, and play, there are very real threats associated with the increased use of technology and our growing dependence on cyberspace.

Individuals and organizations alike depend on cyberspace. From the casual consumer using social media, to online merchants growing their business, to physicians supporting their patients – every sector of our society is increasingly dependent upon technology and networked systems. Without sufficient awareness of the risks in cyberspace, however, behavioral decisions and unseen threats can negatively impact the security of the global cyberspace infrastructure and can cause physical damage in the real world. On an individual level, what is at stake is the vulnerability of each individual user in cyberspace. An individual who is not aware of, and does not implement, basic cyber security practices faces greater personal risk on and offline, such as identity theft, when engaging in daily online tasks. Much of the general public may not be fully aware of the risks of operating in cyberspace, which can affect both personal and national security. All individuals who go online share responsibility for a more secure cyberspace and need to be aware of the risks and vulnerabilities in cyberspace. Protecting personal information and that of others, reinforcing their systems against attacks, and guarding against the use of their own systems in attacks, help individuals create a more secure cyberspace for everyone. Through the Secure Chat System, it will counter these risks and help make cyberspace, especially in instant messenger (IM) [5] more secure. Secure Chat can prepare the general public to identify and avoid risks in instant messenger.

Phishing attack is the major problem in cyber space. In cyber space, especially in instant messenger much of our personal and sensitive information are disclosed through text and voice messages. Detection of phishing attack in instant messenger are now available like Anti phishing Detection APD [1] algorithms, but it is not yet a secure one because the user need to share some personal and sensitive information to identify whether the chat mate is a phisher or not. Here propose an efficient phishing detection system in instant messenger which don't need user personal and sensitive information. The system identify each user behavior according to their chat history and provide appropriate warning messages to users about the chat mate who act as a trust worthy. And in addition to this one the system also provide appropriate messages to user about the behavior of their all chat mate. The proposed work carry out using Data Mining technique with speech recognition system. Words are recognized from speech with the help of FFT spectrum analysis and LPC coefficient methods [17]. The major advantages of this secure chat system is that it is a zero day phishing. It can identify the phishing person without N days of transaction.

II. RELATED WORK

We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and database management systems (DBMS). The efficient database management systems have

been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in data With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

Popular system are use to identify the phishing web sites by the content based approach like CANTINA [2]. It work as collect the most frequent words from a website by using the TF-IDF algorithm and provide the top five result to the popular search engines and compare the url's displayed with the web site and identify the phishing site or not. similarly identifying the common string in phishing web sites and thus help to inform the web server for the common attack methods and types [4]. It focus on studying the structure of URLs employed in various phishing attacks. They find that it is often possible to tell whether or not a URL belongs to a phishing attack without requiring any knowledge of the corresponding page data. They describe several features that can be used to distinguish a phishing URL from a benign one. These features are used to model a logistic regression filter that is efficient and has a high accuracy. They use this filter to perform thorough measurements on several million URLs and quantify the prevalence of phishing on the Internet today.

The paper is organized as section 3 discuss about Secure Chat algorithm, section 4 discuss about some experimental results and section 5 discuss about conclusion and discuss.

III. PROPOSED WORK

This paper propose a new architecture for detecting a misleading phishing attacks in instant messenger. The proposed architecture shows. The proposed work have organized with different stages. Each stage is carry out independently. The different stages are Row data collection, Extract the transaction data, Extract frequent patterns from transaction data, Extract phishing words, Provide user notifications. The databases include in this secure chat is voice database for storing voice data while chatting similarly text database for storing text data. We have an ignore word database which have all the ignore words like is , was , there etc..Similarly there is phishing database which have all the phishing word and shortcut database which have all the short cut words like pwd for password, uname for user name etc.

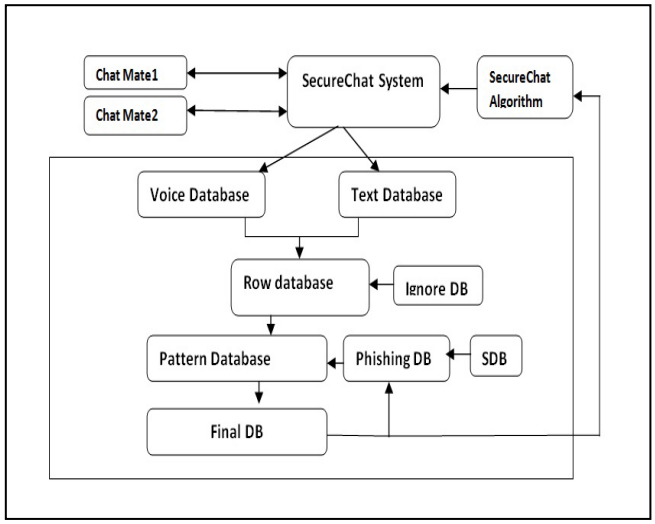


Fig 3.1: System Architecture.

A. Row Data Collection

The first stage of Secure Chat system is to collect the row data. The row data includes voice data and text data during the chatting session of chat mate in instant messenger. The major advantages are here is that Secure Chat don't need N days of transaction in between the chat mates. It collect the row data at the beginning of chatting session. This is carried out by each user have a unique user id at the user registration table, and all the chatting history data stored at table Chat Data. While a new request for chatting has come to the user, the system collect the unique user ids of both chat mate and the system collect the entire chatting data regarding to the new guest chat mate. The Secure Chat system collect all voice and word data corresponding to the guest chat mate. Here the words are recognized from speech with the help of FFT spectrum analysis and LPC coefficient methods. And these word and text data are combined together to form the Row Data Base(RDB).

B. Extract Transaction Data

In all chatting session we have to use many English grammatical words like is, was, there, this etc..We have to ignore all these kind of words. In second stage of Secure Chat system it collect all the data from RDB and remove ignore words from RDB by compare it with ignore word database(IDB). Thus we extract the Transaction DB from IDB and RDB data. The ignore data base have the information of all the ignore words. We collect all the ignore words from this data base and apply it with row data base, which have all the chatting session words like voice and text data.

C. Extract Frequent Patterns

Pattern extraction is the core stage of Secure Chat system. Here we have the Raw data base(RDB) which have all the data

that corresponding to our guest chat mate. Apply Apriory algorithm here to get the frequent patterns from the transaction database. We have to put the support and confidence values. Finally we extract all the patterns that have the minimum support and confidence value. After getting the frequent patterns from transaction data base we store these patterns in to patterns data base. Pattern data base have all the frequent patterns and they do not contain any grammatical words.

D. Extract Phishing Words

Here we apply the Secure Chat algorithm to get the result that the guest chat mate is a phishing person or not. We have the frequent patterns from the transaction database and compare the frequent patterns with the phishing word database (PDB) and short cut word (SDB). If there is a match found in this comparison we have to conclude that the guest chat person tries to get the sensitive information regarding to us. At first we collect all the patterns from the phishing word data base(PDB) and the short cut word data base(SDB). Extract all the frequent patterns from the pattern data base and compare if there is any match found in between this result with previously collected data from phishing word data base and short cut word data base. We extract all the words from here if there is any patterns are match and we have to maintain a value which have all the matching value. If the match count grater than a particular user defined value we have the assumption that the guest mate is phishing one. If there is any new patterns are found we store this new patterns into the phishing word data base.

E. Provide User Notifications

The final stage of Secure Chat system is to provide the user notification . Here after getting a particular threshold value the Secure Chat server provide appropriate user notifications are provided to users about their guest chat mate in order to keep their user account and personal data secure. According to this user notification the user can decide whether to continue the chatting session with his guest. Similarly the user can block this facility according to the system provide the facility.

Secure Chat Algorithm

Input

Voice and Text data.

Output

Notification messages to user according to the prediction.

Steps

1. Initiate Chatting Session
2. New chatting request from guest chat mate
3. Collect all chat history data regarding the guest chat mate from VDB and TDB
4. Ignore all unwanted words from the collected data using IDB
5. Store Resulted data at RDB(Raw Data Base)
6. do

6.1 scan Raw data base

6.2 Apply Apriory Algorithm to collect frequent patterns from Raw data base

6.3 store frequent patterns at pattern database(PDB) until raw data base null

7. do Call Apriory algorithm to scan pattern data base

7.1 calculate confidence and support value of each pattern

7.2 compare the value with phishing DB threshold

7.3 check reach the particular threshold
if yes count the total match and compare with some user defined value

7.4 provide appropriate user friendly notifications

7.5 push the new pattern to Phishing DB

8. repeat until the RDB null

IV. RESULT AND OBSERVATION

The major advantage of this work is that Secure Chat provide appropriate user friendly notifications about the guest chat mate who chat with us. The system will generate these messages automatically by analyzing the chat history date of the guest chat mate. The user can decide whether to continue the chat or not according to the user notifications. The major advantage of this Secure Chat system that it provide security for users sensational data and user privacy. Another advantage is that user need not to share or the system does not need the user chat data. The system only focus on the chat data of the guest chat mate. So we can say that it is a Zero day phishing detection system. Similarly it take the advantages of shortcut words while chatting. The system keep a database which have all the shortcut words to there. Similarly the system can also collect the entire information that are shared by the user to his chat mates. This will help to prevent phishing attacks in shadow stages.

The algorithm Secure Chat is a complex algorithm. It have different stages. The first stage dataset of Secure chat Algorithm is shows in figure. Here the voice and word data are combined and form the chat data. It have user from id, user to id , the chat data and the date of chatting is done

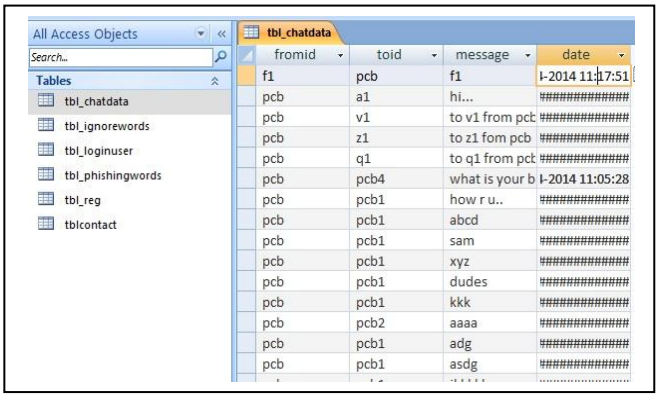


Fig 4.1 : Raw data

The phishing word are stored at phishing database and the figure below show the database. It contain all the phishing word that are expected in chatting session similarly it store the new phishing words patterns that are coming from the chatting session

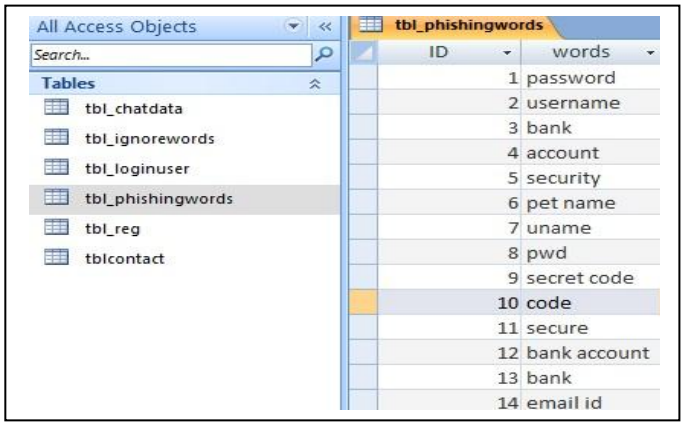


Fig 4.2 : Phishing word data

Similarly we have the following figure shows the phishing word patterns that are extracted using this secure chat algorithm

User Id	Phishing word
pcb1	password
pcb1	username
pcb1	bank
pcb1	bank account
pcb1	bank account username
anil	security code
ajith	secret code
ajith	username password
anil	petname

Fig 4.3 : Phishing words

Precision, recall and f-value are the matrices used for reporting the performance of algorithms. Precision is the ratio of the number of correctly retrieved patterns to the total number of patterns retrieved . Recall is the number of correctly retrieved patterns to the total number of standard patterns. F-value gives the ratio of twice the product of precision and recall to their sum.

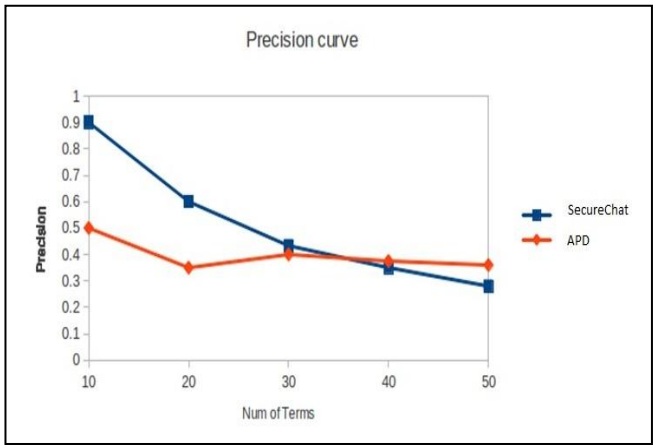
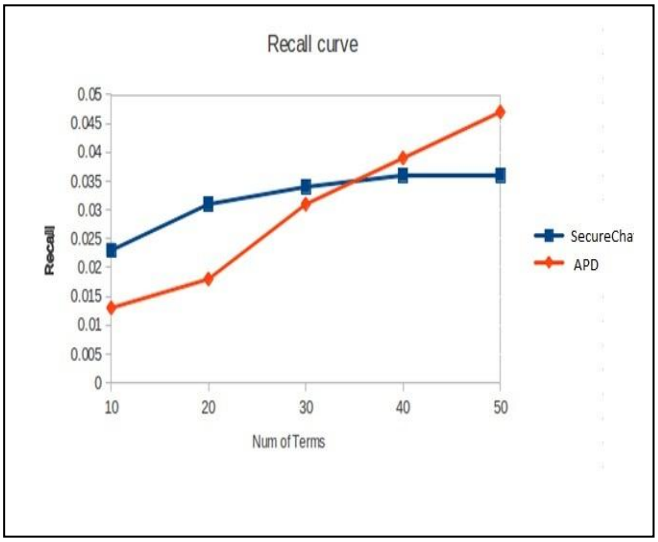


Fig 4.4 precision

From fig 4.4 it can be seen that precision values are higher for Secure Chat and as the number of terms increases a significant improvement can be seen for APD based feature identification. As far as recall is concerned, again Secure Chat shows

Fig 4.5 : Recall

As we consider the recall graph we have better results when compared with APD but its performance begins to degrade when the number of terms increases and APD seems to show good recall values at this point . The graph of F -



value seems to show similar results as to that of recall curve. Therefore it is concluded that as far as Secure Chat dataset is concerned Secure Chat based feature identification gives better results than APD based feature identification.

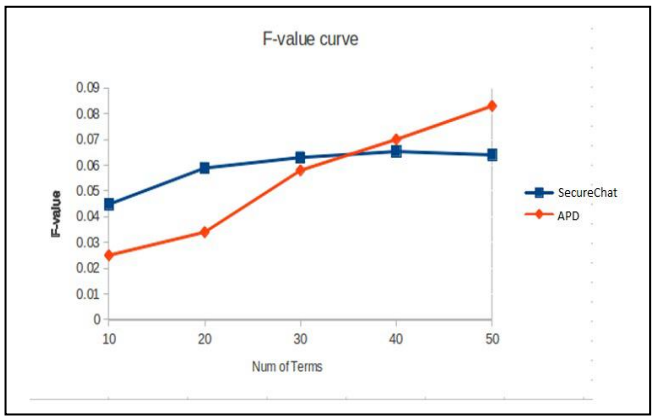


Fig 4.6 : F- Value

V. CONCLUSION AND FUTURE WORK

A data mining approach for Secure chat in IM is implemented. We develop a Secure Chat algorithm based on Apriori algorithm is implemented in this work. The experimental results for this work shows that this one is the most suitable algorithm in instant messenger to identify the misleading phishing attacks and it help to identify phishing thefts and other harmful persons in cyber space. And help the sensational and secret data of user became secure. The future for this approach is look green. The future work of this Secure Chat include we can add video instant messages in this. Similarly the challenges include are the voice consist of numbers and fractional numbers are the challenging one. Similarly the number consist of double works like 22(two two) and roman words and other symbols are challenges.

REFERENCES

[1] M. Sirajuddin, M. N. P. Kumar, M. R. Divya, and M. Rasheed, "Data mining approach for deceptive phishing detection system." Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012

[2] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in Proceedings of the 16th international conference on World Wide Web. ACM, 2007, 639648.

[3] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabatah, "Modelling intelligent phishing detection system for e-banking using fuzzy data mining," in CyberWorlds, 2009. CW'09. International Conference on. IEEE, 2009, pp. 265272.

[4] B. Wardman, G. Shukla, and G. Warner, "Identifying vulnerable websites by analysis of common strings in phishing urls," in eCrime Researchers Summit, 2009. eCRIME'09. IEEE, 2009, pp. 1-13.

[5] R. B. Jennings, E. M. Nahum, D. P. Olshefski, D. Saha, Z.-Y. Shae, and C. Waters, "A study of internet instant messaging and chat protocols," Network, IEEE, vol. 20, no. 4, pp. 16-21, 2006.

[6] M. Atighetchi and P. P. Pal, "Attribute-based prevention of phishing attacks." In NCA, 2009, pp. 266269.

[7] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," Communications of the ACM, vol. 50, no. 10, pp. 94-100, 2007.

[8] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in Proceedings of the 2007 ACM workshop on Recurring malcode. ACM, 2007, pp. 1-8.

[9] A. Herzberg and A. Gbara, "Trustbar: Protecting (even naive) web users from spoo_ng and phishing attacks," Computer Science Department Bar Ilan University, vol. 6, 2004.

[10] R. Segal, J. Cford, J. O. Kephart, and B. Leiba, "Spamguru: An enterprise anti-spam filtering system." in CEAS, 2004.

[11] M. Bishop and D. Frincke, "Joining the security education community," IEEE security & privacy, vol. 2, no. 5, pp. 61-63, 2004.

[12] F. S. Tsai and K. L. Chan, "Detecting cyber security threats in weblogs using probabilistic models," in Intelligence and Security Informatics. Springer, 2007, pp. 46-57.

[13] R. J. Harknett and J. A. Stever, "The new policy world of cybersecurity," Public Administration Review, vol. 71, no. 3, pp. 455-460, 2011.

[14] M. Collins, D. Schweitzer, and D. Massey, "Canvas: a regional assessment exercise for teaching security concepts," in Proceedings from the 12th Colloquium for Information Systems Security Education, 2008.

[15] E. McDuffe, "Nice: National initiative for cybersecurity education," in Proceedings of the Seventh Annual Workshop on Cyber Security and Information Intelligence

[16] Gurpreet singh, "word recognition from speech signal using spectrum analysis and LPC," thesis submitted at thapar university in 2011.