International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

The Detection of Suspicious Email Based on Decision Tree

Samruddhi Rane [1], Gargi Moholkar [2], Kajal Yerunkar [3], Prof. Nilima Nikam [4]

[1,2,3] B.E. Students, Department of Computer Engineering, Yadavrao Tasgaonkar Institute Of Engineering And Technology, Bhivpuri Road, Karjat,
[4] Professor, Department of Computer Engineering, Yadavrao Tasgaonkar Institute Of Engineering And Technology, Bhivpuri Road, Karjat,
Mumbai University, Maharashtra,India

Abstract - *There are many places where we have local area networks and many people using them as per their own needs. In such conditions, we have to closely monitor the computers. Many a times, there comes a situation when we need to lock the resources such as drives, folders or files on these computers to restrict the users of making use of them. Sometimes we need to stop the users from using the internet or from changing the setting or accessing the registry editor so as to secure the systems from any crash due to misuse of it. These are the common task that we do in our day to day life but for this we don't have utility software. The aim of this project is to suspect the E-mails which consist of offensive, anti-social elements and block them which help in identifying the suspicious user. In this project, suspicious users are identified by determining the keywords used by him/her. The keywords such as bomb, RDX, are found in the mails which are sent by the user. All these blocked mails are checked by the administrator and identify the users who sent this mail. This is very useful in real-time scenario in which you can resume the Criminal activities.*

Key Words: secure the systems, suspect the E-mails, block, suspicious users, Criminal activities

1. INTRODUCTION

The idea behind selecting in-house project is to develop software which gives wider scope to express our knowledge. Email has been an efficient and popular ,work done by various researches suggests that deceptive writing is characterized by reduced frequency of first person pronouns and exclusive words and elevated frequency of negative emotion words and action verbs. In many security informatics applications it is frequency of first-person pronouns and exclusive important to detect deceptive communication in email. We apply this model of deception to the set of Email .The rules generated are used to test the email as deceptive or not. In particular we are interested in detecting emails about criminal activities. After classification we must be able to differentiate the emails giving information about past criminal activities (Informative email) and those acting as alerts (warnings) for the future criminal activities. This differentiation is done using the features considering the tense used in the emails. Experimental results show that simple associative classifier provides promising detection rates.

Our purpose is to develop software that can be used:

1. To find a system that identifies deception in Email through communication.
2. Classified deceptive Email.
3. After classification of deceptive Email, we must classify them into informative Emails and alert Emails.

Concern about National security has increased since the terrorist attack on September 2001.The CIA, FBI and other federal agencies are actively collecting domestic and foreign intelligence to prevent future attacks. These efforts have in turn motivated us to collect data's and undertake this paper work as a challenge. Data mining is a powerful tool that enables criminal investigators who may lack extensive training as data analyst to explore large databases quickly and efficiently. Computers can process thousands of instructions in seconds, saving precious time. In addition, installing and running software often costs less than hiring and training personality. To our knowledge, the first attempt to apply Association rule mining to task of suspicious Email Detection (Emails about criminal activities). The reason they have eluded conclusion extracting the informative emails using the tense (Past tense) of the verbs used in the emails. Apart from the informative emails, other emails are considered as the alerting emails for the future occurrences of hazardous activities.

Objective is to protect the security and morale of nation which is being broken down due to the constant attack and threat on the people of our nation. This project helps us to detect the emails sent by the people on internet which can harm out integrity. By the use of this technique, the emails can be found out in a manner which can inform us about past activities as well as alert us about the future activities. Hence it can be used in cyber investigation for the security from suspicious emails.

1.1 Existing System

Aniruddha Kshirsagar and Hanmant N. Renushe proposed that various data mining techniques can be used to overcome the critical issue of identity fraud. They proposed

that data mining can be used in almost every field to detect the frauds and crimes in various domains.

Few other authors have also applied ID3 algorithm to detect the suspicious e- mails by considering the keyword base approach.

Disadvantages:

- No proper identification of deceptive emails.
- The problem with this approach was that they have not used any information about the context of the identified keywords from the e-mails.

1.2 Proposed System

The proposed system is developed in Microsoft Visual Studio 2010 using ID3 Classification Algorithm. At the back end MySQL Server has been used and an input file has been created for the training sample set of e-mails. Some rules are provided by the new Decision Tree which further are used to detect that a particular e-mail is Suspicious or not. This application is platform independent. It is an on click application which once implemented through Visual Studio does not require any kind of software. The functioning is very simple and user friendly.

The proposed system initially extracts some useful features such as "suspicious keywords" and "non-suspicious indicators" from the e-mail message. Then the combination of those keywords and indicators are analyzed. If suspicious keywords are present in an e-mail without any non-suspicious indicator, the e-mail will be detected as suspicious and the threat of a potential future terrorist event will be reflected. If some suspicious keywords are present with others i.e., non-suspicious indicators, then the e- mail will be further classified as "may-be- suspicious" because it might be the case of an e-mail in which people discusses past events, maybe for condolence, sympathy and so on. In this tool, the features will be extracted along with the context.

For example, take a sample e-mail such as: "It is very sad that you couldn't save your family from the bomb blasts." In the e-mail message, the suspicious keywords "blasts" and "bomb" are present but they do not reflects a future terrorist activity due to the presence of another feature "sad".

ID3 Decision tree algorithm has been used to classify the records. Algorithm starts with a training set T = {mail1, mail2, mail3…mail n} and class label is Suspicious = {Yes, No, Maybe}. Each e-mail is provided a label. The purpose is to develop a tool that observes the patterns from the training sample set and is able to classify a new test sample e-mail as suspicious, non-suspicious or may-be-suspicious. Algorithm extracts the Suspicious Keywords and the Non-Suspicious Indicators which were present in the training sample set of e-mails.

While constructing the Tree, at start, all the suspicious keywords are at the root. Then those are partitioned recursively based on selected attributes which are selected on the basis of the information gain after attribute- importance. The new algorithm through introducing attribute -importance emphasizes on the attributes with less values but higher importance, rather than the attributes with more values and lower importance.

Further partitioning is stopped when there are no remaining attributes for further partitioning i.e. majority voting is employed for classifying the leaf or there are no samples left.

For all the attributes, the attribute- importance is calculated. The attribute which has the highest importance becomes the root node of the tree. This process goes on until all the attributes are mapped in to the tree based on the sorted attribute - importance. Following the each individual path in the tree, the rules are generated. The output of this module is Decision tree and Rules.

1.3 E- MAIL

Electronic mail, also known as email is a method of exchanging digital messages from an author to one or more recipients. Modern email operates across the Internet or other computer networks. Some early email systems required that the author and the recipient both be online at the same time, in common with instant messaging. Today's email systems are based on a store-and-forward model. Email servers accept, forward, deliver and store messages. Neither the users nor their computers are required to be online simultaneously; they need connect only briefly, typically to an email server, for as long as it takes to send or receive messages.

An Internet email messages consists of three components, the message envelope, the message header, and the message body. The message header contains control information, including, minimally, an originator's email address and one or more recipient addresses. Usually descriptive information is also added, such as a subject header field and a message submission date/time stamp.

1.4 Read Email Using IMAP

Internet Message Access Protocol (IMAP) is one of the two most prevalent Internet standard protocols for e-mail retrieval, the other being the Post Office Protocol (POP). Virtually all modern e-mail clients and mail servers support both protocols as a means of transferring e-mail messages from a server.

The Internet Message Access Protocol (commonly known as IMAP) is an Application Layer Internet protocol that allows an e-mail client to access e-mail on a remote mail server. The current version, IMAP version 4 revision 1 (IMAP4rev1), is defined by RFC 3501. An IMAP server typically listens on well-known port 143. IMAP over SSL (IMAPS) is assigned well-known port number 993. IMAP supports both on-line and off-line modes of operation. E-mail clients using IMAP generally leave messages on the server until the user explicitly deletes them. This

and other characteristics of IMAP operation allow multiple clients to manage the same mailbox. Most e-mail clients support IMAP in addition to POP to retrieve messages; however, fewer e-mail services support IMAP. IMAP offers access to the mail storage.ss Incoming e-mail messages are sent to an e-mail server that stores messages in the recipient's e-mail box. The user retrieves the messages with an e-mail client that uses one of a number of e-mail retrieval protocols. Some clients and servers preferentially use vendor-specific, proprietary protocols, but most support the Internet standard protocols, SMTP for sending e-mail and POP and IMAP for retrieving e-mail, allowing interoperability with other servers and clients. For example, Microsoft's Outlook client uses a proprietary protocol to communicate with a Microsoft Exchange Server server as does IBM's Notes client when communicating with a Domino server, but all of these products also support POP, IMAP, and outgoing SMTP. Support for the Internet standard protocols allows many e-mail clients such as Pegasus Mail or Mozilla Thunderbird to access these servers, and allows the clients to be used with other servers (see list of mail servers).

2. ID3 ALGORITHM

The ID3 algorithm begins with the original set ***S*** as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set ***S*** and calculates the entropy ***H(S)*** (or information gain ***IG(S)***) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set ***S*** is then split by the selected attribute to produce subsets of the data. The algorithm continues to recurs on each subset, considering only attributes never selected before.
Recursion on a subset may stop in one of these cases:

- every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labelled with the class of the examples
- there are no more attributes to be selected, but the examples still do not belong to the same class (some are + and some are -), then the node is turned into a leaf and labelled with the most common class of the examples in the subset
- there are no examples in the subset, this happens when no example in the parent set was found to be matching a specific value of the selected attribute

Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.
The main process of id3 algorithm is:

- Calculate the entropy of every attribute using the data set ***S***
- Split the set ***S*** into subsets using the attribute for which the resulting entropy (after splitting) is minimum (or equivalently, information gain is maximum)
- Make a decision tree node containing that attribute
- Recurs on subsets using remaining attributes

3. LITERATURE SURVEY

Aniruddha Kshirsagar et al. [3] and Hanmant N. Renushe et al. [4] proposed that various data mining techniques can be used to overcome the critical issue of identity fraud. They proposed that data mining can be used in almost every field to detect the frauds and crimes in various domains. P. S. Kaila and Skillicon [8] proposed an approach for detecting the suspicious and deceptive e-mails which is dependent on the singular value decomposition method. The problem with this approach was that it does not deal with incomplete data in a proper efficient way and cannot maintain the new data incrementally. For updating the new data properly they had to reprocess the entire algorithm. S. Kiritchenko et al. [9] projected a good performance minimizing the Classification error in an e-mail sequence by observing temporal relations and embedding the discovered information into content-based learning methods. Approach to Anomalous e-mail detection is considered. Z. Huang et al. [12] showed approaches to detect anomalous e-mail and involved the deployment of data mining techniques.
Few other authors [6] have also applied ID3 algorithm to detect the suspicious e- mails by considering the keyword base approach. The problem with this approach was that they have not used any information about the context of the identified keywords from the e-mails. Then S. Appavu & R. Rajaram [10] have used the association rule mining to further classify the suspicious e-mails into different specialized classes. This system decides whether the e-mail message can be classified as suspicious alert if there is suspicious keyword present in the future tense or else it is classified as suspicious information only.

4. SYSTEM REQUIREMENT

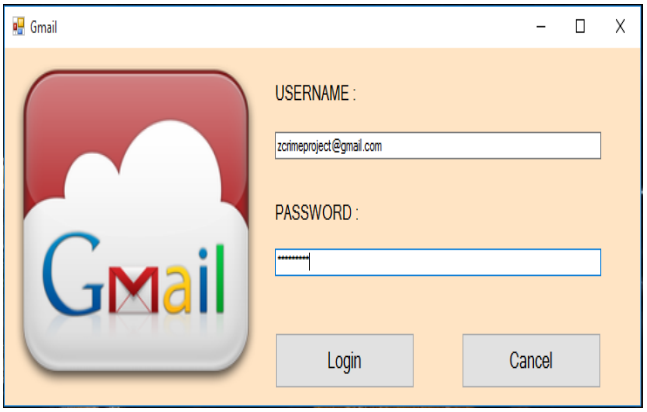
4.1 Software Requirements

- O/S : Windows XP.
- Language : .net, C#
- IDE : Microsoft Visual Studio 2010
- Data Base : MySQL

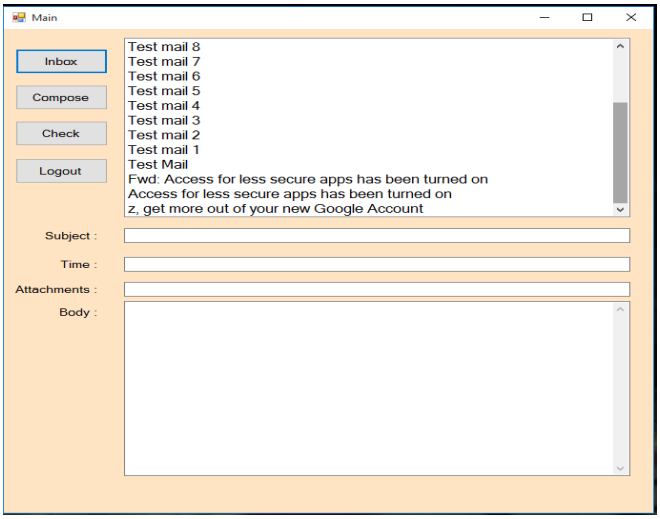
4.2 Hardware Requirements

- System : Pentium IV 2.4 GHz
- Hard Disk : 16 GB
- Ram : 512MB

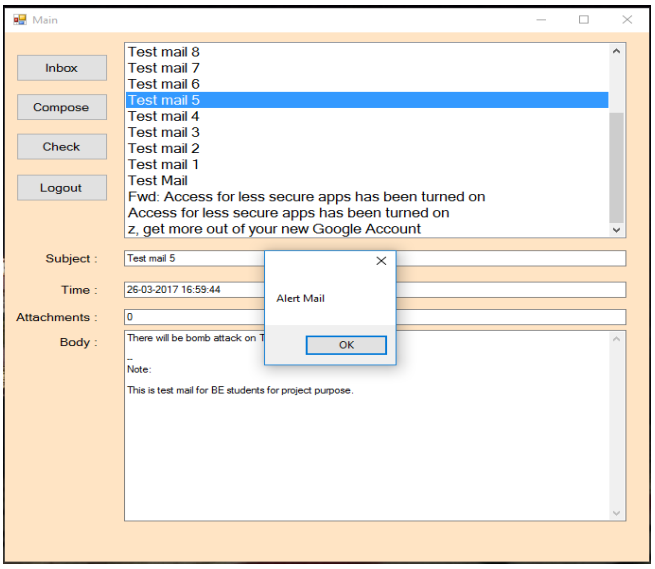
5. SCREENSHOTS



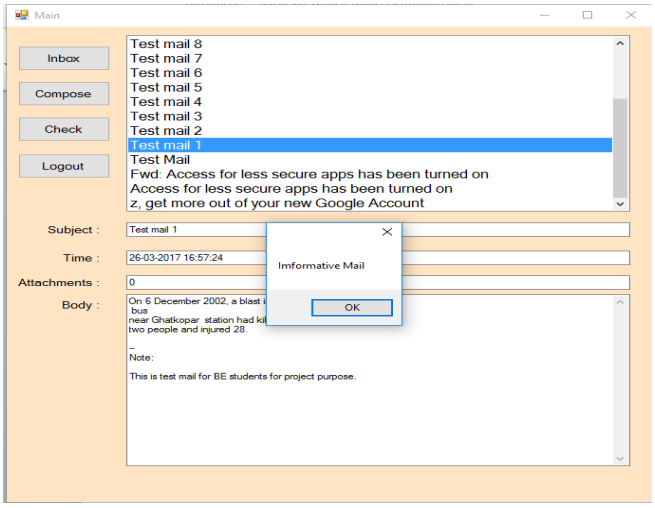
Screenshot - 1: Users login



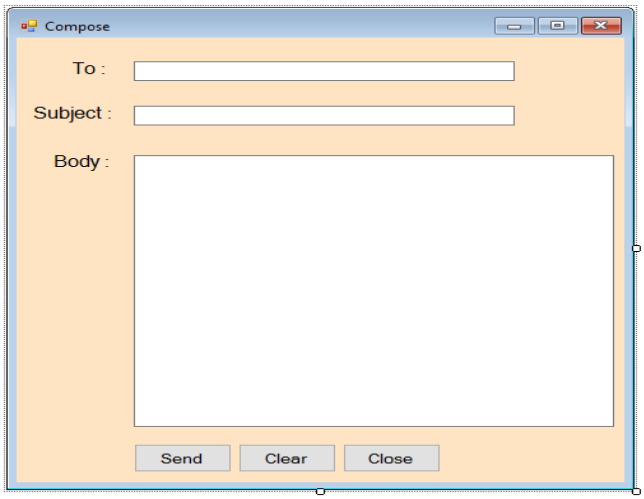
Screenshot - 2: Inbox



Screenshot – 3: Check Mail 1 (Alert Mail)



Screenshot - 4: Check Mail 2 (Informative Mail)



Screenshot - 5: Compose New Mail

6. APPLICATIONS

Our developed software can be used:

- In crime branch to detect the suspicious mails & to get alert of several attacks.
- In Army, to give alert of attacks & blast.
- To provide security against cyber crime.

7. CONCLUSION

We have deployed decision tree based classification approach to detect deceptive communication in email text as informative or alert emails. We can find it that the ID3 algorithm can provides better classification result for suspicious email detection.

The Proposed work (using keyword extraction and keyword attribute called tense) will be helpful for identifying the deceptive email and to get information in time to take effective actions to reduce criminal activities.

REFERENCES

[1] Mugdha Sharma, "Z - CRIME: A data mining tool for the detection of suspicious criminal activities based on decision tree", International Conference on Data Mining and Intelligent Computing (ICDMIC), 2014.

[2] Keyvanpour, Javideh, et al., "Detecting and investigating crime by means of data mining: a general crime matching framework", World Conference on Information Technology, Elsvier B.V., 2010.

[3] Aniruddha Kshirsagar & Lalit Dole , "A review on data mining methods for identity crime detection", International Journal of Electrical, Electronics and Computer Systems (IJEECS), January, 2014.

[4] N. Hanmant Renushe, R. Prasanna Rasal & S. Abhijit Desai, "Data mining practices for effective investigation of crime", IJCTA, May-June 2012.

[5] T. Abraham and de Vel, O., "Investigative profiling with computer forensic log data and association rules", Data Mining, Proceedings, IEEE International Conference, 2002.

[6] S. Appavvu, Muthu Pandian, et al., "Association Rule Mining for Suspicious Email Detection: A Data Mining Approach", Intelligence and Security Informatics, IEEE, 2007.

[7] M. Mansourvar, et al., "A computer- Based System to Support Intelligent Forensic Study", Computational Intelligence, Modelling and Simulation (CIMSiM), Fourth International Conference, IEEE, September 2012.

[8] P.S.Keila, D.B. Skillicorn, "Detecting Unusual and Deceptive Communication in Email", Centers for Advanced Studies Conference, June 2005.

[9] S. Kiritchenko, S. Matwin, S. Abu-Hakima, "Email Classification with Temporal Features", Intelligent Information Systems, 2004.

[10] S. Appavu, R. Rajaram, "Association rule mining for suspicious email detection: a data mining approach", In Proceedings of the IEEE International Conference on Intelligence and Security Informatics, New Jersey, USA, 2007.

[11] Sarwat Nizamani, Nasrullah Memon, et al., "Modeling Suspicious Email Detection using Enhanced Feature Selection", International Journal of Modeling and Optimization, Vol. 2, No. 4, August 2012.

[12] Z.Huang, D.D. Zeng, "A Link Prediction Approach to Anomalous Email Detection", Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan, 2006.