Mining E-mail Content for Cyber Forensic Investigation

Ms. Sobiya R. Khan
P.G. Dept. of Computer Sci.
&Eng., GHRCE, Nagpur, India
sobi16@gmail.com

Ms. Smita M. Nirkhi
P.G. Dept. of Computer Sci.
&Eng.,GHRCE, Nagpur, India
smita811@gmail.com

Dr. R. V. Dharaskar
M. P. G. I,
Nanded, India
rvdharaskar@rediffmail.com

***Abstract* - E-mail is a widely used mechanism for communication, due to its cost and expediency. However, the concern lies when along with its legitimate usage; it is being abused for committing various cyber crimes. E-mail system security lacks adequate proactive mechanism, to defend against such vulnerabilities and misuses. A cyber forensic investigation is employed for gathering significant evidences against adversaries by examining suspected e-mail accounts, in order to prosecute criminals in court of law. In this context, data mining techniques and tools based on them have been used extensively for extracting evidences from huge e-mail ensembles. This can provide assistance to the forensic investigator, to perform a multi-staged analysis of e-mail ensembles. In this paper, we briefly discuss various applications of data mining techniques with respect to cyber forensic investigation. Specifically, we describe our proposed framework and give implementation of first module,e-mail statistical analysis of our framework.**

***Keywords* - Cyber Crime, E-mail forensic analysis, Statistical Analysis, Classification and Clustering techniques, Authorship identification, Community identification.**

I. INTRODUCTION

Nowadays, e-mail has become an easy, efficient and economical means of communication over the Internet & Intranet. It is being employed by most of the industries and governments, as well. Thus there is huge amount of e-mail traffic generated on daily basis. However, with its increased usage, there is an undesired increase in the crimes which are mediated via e-mails. Examples of such misuse include: phishing, spamming, drug trafficking, cyber bullying, racial vilification, child pornography, and sexual harassment. The prime reason for this inherent vulnerability are twofold, firstly, there is no mechanism for message encryption at the sender end and an integrity check at the recipient end. Secondly, the widely used, Simple Mail Transfer Protocol (SMTP) e-mail protocol lacks a source authentication mechanism. This inherent vulnerability of e-mail communication exposes it to such crimes. Due to such crimes, these e-mail misuse phenomena do a lot of harm to people‘s benefit, and even influence social stability.However, there are no effective upbeat methods for preventing these phenomena. The current methods are merely some passive mechanisms such as e-mail filtering, installing firewall, etc. But they are unable to put an end to the e-mail misuse phenomena.In this context, a cyber crime investigation is carried out to gather evidences and bring the culprits in the court of law and provide justice to the victims. This had lead to the need for efficient automated tools in the hands of forensic experts, during forensic investigation, which can provide means to capture evidence against such criminals which are credible in the court of law.

This paper is divided into six sections. Section II briefly describes the issues to be considered in e-mail mining. Section III gives a brief overview of the related work. Section IV gives an outline of our proposed framework, its experimental setup and implementation of first module e-mail statistical analysis of our framework. Theexperimentalresults are described in Section V. Section VI gives the conclusion drawn.

II. E-MAIL MINING

Digital Forensic technology has already become a centre of attention among researcher’s and law professionals. With the ongoing research and development in this technology, there is hope that this could help to curb the amount of cyber crime going.Data mining is the process of extracting useful patterns from vast amount of data. So, the obvious question is why to use data mining in forensic investigation? The crux to this question lies in the Analysis phase of forensic investigation. The Analysis phase poses difficulty in front of the forensic investigator, because it’s difficult to analyse large data set, if no appropriate methods are available to process it. Also, it is unknown at the initial stage of the investigation, which pieces of information may have value as evidence. Data mining techniques are inherently applicable to this problem domain and hence can provide an efficient mechanism to capture relevant information out of huge data set [24].

E-mail Mining can be considered as an application of the upcoming research area of Text Mining on e-mail data [17]. However, there are some specific characteristics of e-mail data that set a distinctive separating line between E-mail and Text Mining:

1. Information in the headers of e-mail can be used for various e-mail mining tasks.Text mining techniques might be inefficient fore-mail, as e-mail data is generally quite short.
2. Well-formed linguistic is not guaranteed in e-mail and spelling and grammar mistakes might also appear frequently.Different topics may be discussed; which makes e-mail classification more difficult.
3. E-mail is writtenpersonally hence generic techniques are difficult to be effective.Concepts or distributions of target classes may change over time. HTML tags and

attachments must also be removed in order to apply a text mining technique.

4. Due to privacy issues very few e-mail data are available publicly for experiments. Exception to the above statement, are the Enron Corpus and the Ling Spam corpus, which have been made public for research purpose [24].

III. RELATED WORK

A. E-mail Analysis Tools

Researchers have employed existing state-of-the-art data mining techniques, machine learning algorithms and visualization techniquesto implement various tools and frameworks. Such tools have varied functionalities and applications with respect to cyber crime investigation. Some of the existing tools are online such as e.g., MET, UnMask, etc which use online e-mail data whereas others are offline in nature, such as e.g., EMT, IEFAF, etc which use offline E-mail dumps to extract information relevant for forensic investigation [24]. The following is a brief description of the various tools: The benchmark tools of E-mail Mining Toolkit (EMT) and Malicious E-mail Tracking (MET) were developed at the Columbia University, which employed data mining techniques to perform behaviour-based analyses and social network analysis [8-9]. The EMS toolkit sheds some light on the social network of the users [3]. Visualize Association inside E-mails (VAIE) builds data models of e-mails to classify them in different categories based on key word search techniques[11].Visualization techniques have been employed for e-mail analysis to provide graphical representation of the e-mail data [10].UnMask, has been developed as an ongoing project for determining phishing [12]. IEFAF includes features such as computing statistical distribution; generating data mining models& performing e-mail authorship analysis [1].

B. E-mail Author Attribution

Authorship analysis is a process of examining the characteristics of a piece of writing to draw conclusions on its authorship. Its roots are from a linguistic research area called stylometry, which refers to statistical analysis of literary style. Authorship analysis is categorized into three major field, Authorship identification, Similarity detection and Authorship characterization. Authorship identification determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author. Authorship analysis has been used in a small but diverse number of application areas. Examples include identifying authors in literature, in program code, and in forensic analysis for criminal cases. Authorship analysis has been applied to online messages in recent years [24]. Commendable results were obtained with respect to e-mail authorship analysis on both aggregated and across different topics in [6, 14]. In another literature, four types of writing-style features (lexical, syntactic, structural, and content-specific features) along with SVM were used to identify plausible author of e-mail and online messages [4, 5] which was extended using genetic algorithm in [19]. Stylometric features combined with unsupervised techniques have been employed for author identification and similarity detection in [18]. A novel method termed as Write-print using frequent pattern mining has been developed in [2], which was further improved using clustering technique in [7].

C. E-mail Classification and Clustering

Most e-mail mining tasks are being accomplished by using e-mail classification at some point. E-mail classification is the assignment of an email message to one of the category, from a pre-defined set of categories. Automatic email classification aims at building a model (typically by using machine learning techniques), which will undertake this task on behalf of the user. Examples of applications are automatic mail categorization into folders, spam filtering and author identification [24]. Four different classifiers (Neural Network, SVM, Naïve Bayesian and Decision Tree) have been used to identify suspicious mails in [15]. Naïve Bayesian classifier has been used for identifying threats from a company's rapidly expanding e-mail data set in [16]. Various studies have revealed that SVM has been shown to be very robust and successful. Clustering techniques goes one step further where training data set isn’t available by automatically categorizing data. Clustering technique has been used extensively for text categorization and authorship analysis as well [24]. An effective digital text analysis strategy has been given in [20] which rely on clustering based text mining technique.

D. E-mail Social Network Analysis

Social Network analysis is the study of communication links or associations between people. It reveals a great deal of information about his/her behaviour and circle of people (friends, colleagues, family members, etc.) around him/her with whom he/she interacts [24]. Social Network has been explored in [22] by implementing a novel algorithm using data mining to identify user behaviours, identify patterns of communications between entities in an e-mail collection to extract social standing. Associations between members have been extracted to discover criminal communities in [13]. Social Network analysis has also been explored in [23] which use recursive data mining in order to identify frequently occurring communities in online messages such as e-mails, blogs, chats, etc. Studies have shown that frequent pattern mining techniques have been very successful in this problem domain.

IV. PROPOSED SYSTEM

Above discussed tools, frameworks and techniques have shown expertise in one or the other application, but still they lack a consistent interface, an integrated approach, and a commercial outfit which can provide varied functionalities to analyse e-mail ensembles and discern useful information from it, which could be useful in the investigation process. The results should be available on timely basis during the investigation and evidences should be in such form which could be satisfactorily presented in the court of law for further jurisdiction. Thus, we can conclude that need still exist for automated forensic tools whichwill help forensic experts to efficiently analyse e-mail collections, within a limited time frame.

Here we are proposing the implementation ofa framework which will employ data mining techniques to achieve the various functionalities. The framework is proposed to perform E-mail Statistical Analysis, E-mail Classification & Clustering, E-mail Author Identification and E-mail Social Network Analysis and will try to overcome the limitations observed in previous systems.To evaluate our implementation, we are using the Enron e-mail corpus made available by MIT athttp://www.cs.cmu.edu/~enron/. The proposed framework will be implemented in Java and will use the data mining tool weka.

A. E-mail Statistical Analysis

Statistical analysis of e-mail accounts calculated from communication patterns reflects a great deal of information which could be of value to the forensic investigator. The various possible statistics obtained from the email corpus could be number of e-mails per sender, per recipient, per sender domain, per recipient domain, per class, per cluster etc. Statistics related to Classes and clusters are determined after applying classification and clustering respectively. Computing various statistics such as e-mailing frequency during different parts of the day, average e-mail size, and average attachment size (if any) helps to discern usage behaviour and can be used to detect suspicious behaviour. This may help investigators to narrow down the investigation scope by short listing e-mail accounts that are showing unusual behaviourand can be emphasized for further investigation. Additional information like total number of users (senders/recipients) within an e-mail collection, finding all the recipients of each user can help during investigation.Statistical distributions can be computed over a certain period of time and for a specific set of e-mails.

B. E-mail Extraction

The objective of the first module of our proposed framework is to perform E-mail Statistical Analysis. This provides a statistical summary of the e-mail data, via applying various statistical measures on the data obtained from e-mail. But prior to that, it is essential to extract the relevant information from the e-mail dataset and to represent that information in a form which will be suitable for manipulation for statistical analysis. Hence we can put forth the present objective before statistical analysis as,

1. E-mail extraction
2. Data cleaning for removing ambiguities
3. Identification of relevant tokens
4. Extraction of tokens from e-mail archives such as Message Id, To, From, Date, Time, CC, BCC, Subject and Body
5. Storing the extracted tokens in the Database

The tokens are identified from this data, which is saved inside database and which helps to perform various analyses pertaining to E-mail.

C. E-mail Extraction Logic

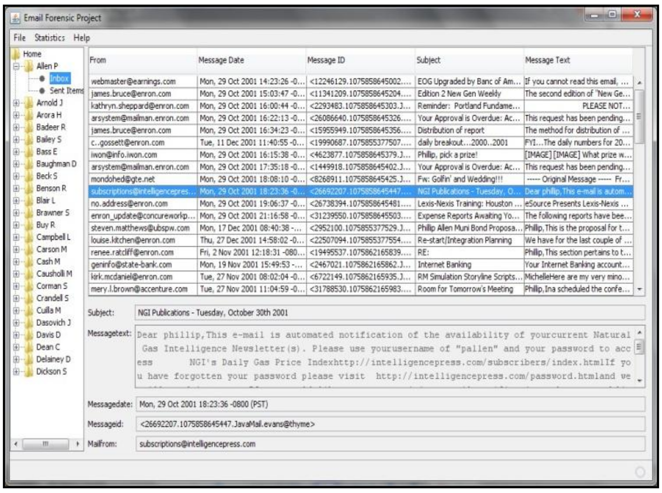
The following steps must be followed for E-mail extraction:

1. Read directory of particular User
2. If file present in directory, Read File
3. For each line in the file, check whether token exists or not. If token is present, extract token value and go to next line
4. Repeat step 3 until end of file
5. If directory has more files go to step 2
6. If all files in directory are processed , stop

After the extraction, various statistics described in the next section are calculated from the database.

V. EXPERIMENTAL RESULT

Presently we are using data from Inbox and Sent folders for our analysis. The E-mail Details Window is as shown in Fig.1. Here the details regarding each users mails is displayed from the inbox and sent folders.



From	Message Date	Message ID	Subject	Message Text
webmaster@earnings.com	Mon, 29 Oct 2001 14:23:26 -0...	<12246129.1075858645002....	EOG Upgraded by Banc of Am...	If you cannot read this email, ...
james.bruce@enron.com	Mon, 29 Oct 2001 15:03:47 -0...	<11341209.1075858645204....	Edition 2 New Gen Weekly	The second edition of "New Ge...
kathryn.sheppard@enron.com	Mon, 29 Oct 2001 16:00:44 -0...	<2293483.1075858645303.J...	Reminder: Portland Fundame...	PLEASE NOT...
arsystem@mailman.enron.com	Mon, 29 Oct 2001 16:22:13 -0...	<26086640.1075858645326....	Your Approval is Overdue: Ac...	This request has been pending...
james.bruce@enron.com	Mon, 29 Oct 2001 16:34:23 -0...	<15955949.1075858645356....	Distribution of report	The method for distribution of ...
c..gossett@enron.com	Tue, 11 Dec 2001 11:40:55 -0...	<19990687.1075855377507....	daily breakout....2000..2001	FYI...The daily numbers for 20...
iwon@info.iwon.com	Mon, 29 Oct 2001 16:15:38 -0...	<4623877.1075858645379.J...	Philip, pick a prize!	[IMAGE] [IMAGE] What prize w...
arsystem@mailman.enron.com	Mon, 29 Oct 2001 17:35:18 -0...	<1449918.1075858645402.J...	Your Approval is Overdue: Ac...	This request has been pending...
mondohed@gte.net	Mon, 29 Oct 2001 18:08:10 -0...	<8268911.1075858645425.J...	Fw: Golfin' and Wedding!!!	----- Original Message ----- Fr...
subscriptions@intelligencepres...	Mon, 29 Oct 2001 18:23:36 -0...	<26692207.1075858645447....	NGI Publications - Tuesday, O...	Dear phillip,This e-mail is autom...
no.address@enron.com	Mon, 29 Oct 2001 19:06:37 -0...	<26738394.1075858645481....	Lexis-Nexis Training: Houston ...	eSource Presents Lexis-Nexis ...
enron_update@concureworkp...	Mon, 29 Oct 2001 21:16:58 -0...	<31239550.1075858645503....	Expense Reports Awaiting Yo...	The following reports have bee...
steven.matthews@ubspw.com	Mon, 17 Dec 2001 08:40:38 -...	<2952100.1075855377529.J...	Philip Allen Muni Bond Proposa...	Philip, This is the proposal for t...
louise.kitchen@enron.com	Thu, 27 Dec 2001 14:58:02 -0...	<22507094.1075855377554....	Re-start/Integration Planning	We have for the last couple of ...
renee.ratcliff@enron.com	Fri, 2 Nov 2001 12:18:31 -080...	<19495537.1075862165839....	RE:	Philip, This section pertains to t...
geninfo@state-bank.com	Mon, 19 Nov 2001 15:49:53 -...	<2467021.1075862165862.J...	Internet Banking	Your Internet Banking account...
kirk.mcdaniel@enron.com	Tue, 27 Nov 2001 08:02:04 -0...	<6722149.1075862165935.J...	RM Simulation Storyline Scripts...	MichelleHere are my very mino...
mery.l.brown@accenture.com	Tue, 27 Nov 2001 11:04:59 -0...	<31788530.1075862165983....	Room for Tomorrow's Meeting	Philip,Ina scheduled the confe...

Subject: NGI Publications - Tuesday, October 30th 2001

Messagetext:
```
Dear phillip,This e-mail is automated notification of the availability of yourcurrent Natural
 Gas Intelligence Newsletter(s). Please use yourusername of "pallen" and your password to acc
ess       NGI's Daily Gas Price Indexhttp://intelligencepress.com/subscribers/index.htmlIf yo
u have forgotten your password please visit  http://intelligencepress.com/password.htmland we
```

Messagedate: Mon, 29 Oct 2001 18:23:36 -0800 (PST)

Messageid: <26692207.1075858645447.JavaMail.evans@thyme>

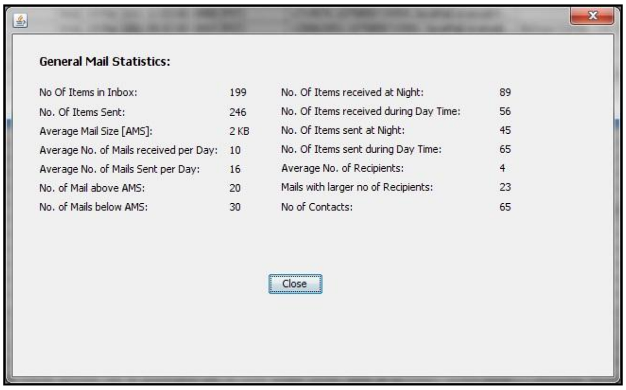
Mailfrom: subscriptions@intelligencepress.com

Figure 1. E-mail Details Window

Fig. 2 represents the General E-mail Statistics which can be calculated from the E-mail data. These statistics are calculated from the E-mail Inbox and Sent data. The various statistics included are as follows:

No. of Items in Inbox
No. of Items Sent
Average Mail Size [AMS]
Average No. of Mails received per Day
Average No. of Mails Sent per Day
Average No. of Mails Sent per Day
No. of Mails below AMS
No. of Items received at Night
No. of Items received during Day Time
No. of Items sent at Night
No. of Items sent during Day Time

Average No. of Recipients
Mails with larger no of Recipients
No of Contacts



General Mail Statistics:

No Of Items in Inbox:	199	No. Of Items received at Night:	89
No. Of Items Sent:	246	No. Of Items received during Day Time:	56
Average Mail Size [AMS]:	2 KB	No. Of Items sent at Night:	45
Average No. of Mails received per Day:	10	No. Of Items sent during Day Time:	65
Average No. of Mails Sent per Day:	16	Average No. of Recipients:	4
No. of Mail above AMS:	20	Mails with larger no of Recipients:	23
No. of Mails below AMS:	30	No of Contacts:	65

Figure 2. E-mail Statistical Analysis

Graphical representation of the statistical data is presented for few statistics in Fig. 3 and 4.

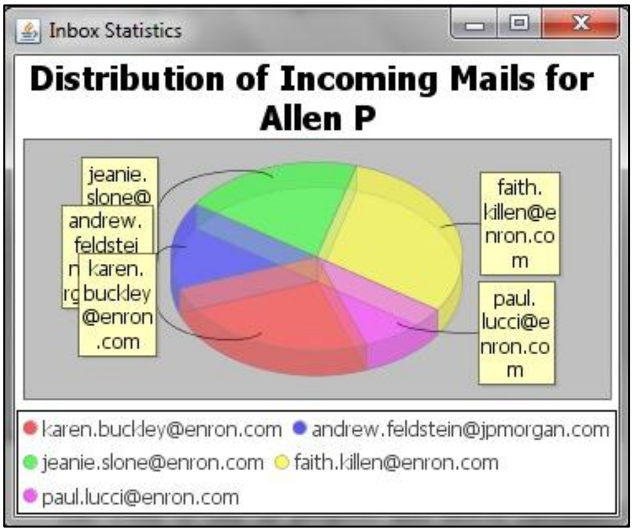


Figure 3. Distribution of Incoming E-mails

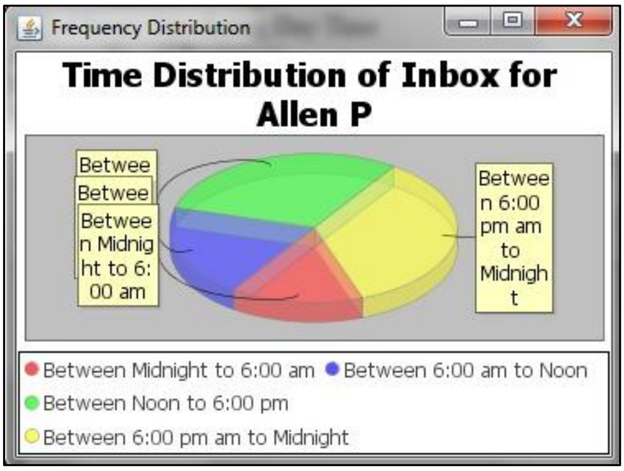


Figure 4. Time Distribution of Inbox

The graph of Fig. 4 represents inbox distribution of 5 users for user Allen P. The next graph in Fig. 5 shows the mailing frequency during various time distributions. Anomalous behaviour can be identified from the statistical distributions and unusual behaviour could be identified by the forensic investigator. The work is still in progress for rest of the modules.More Statistics will be added to the general statistics after the completion of our second and fourth module, which will include the statistics corresponding to Classification, Clustering and Social Network Analysis.

VI. CONCLUSION

In this paper, we briefly discussed the application of data mining techniques with respect to various automated tools, e-mail authorship analysis, e-mail classification & clustering and social network analysis. The study of previous work reveals that data mining techniques gives a promising lookfor analysis of huge e-mail dataset. Automated tools based on such technique can assistforensic investigator during initial cyber forensic investigation. We are employing data mining techniques to implement our framework. The initial results of our first module of E-mail Statistical Analysis have been shown which will be integrated with the rest of the modules.

REFERENCES

[1] Rachid Hadjidj, Mourad Debbabi, Hakim Lounis, Farkhund Iqbal, Adam Szporer, Djamel Benredjem, “Towards an integrated e-mail forensic analysis framework”, Digital Investigation 5, pp.124–137, 2009.

[2] Iqbal F, Hadjidj R, Fung BCM, Debbabi M., “A novel approach of mining write-prints for authorship attribution in e-mail forensics”, Digital Investigation 5:pp.42–51, 2008.

[3] Hongjun Li, Jiangang Zhang, Haibo Wang, Shaoming Huang, “A Mining Algorithm For E-mail’s Relationships Based On Neural Networks”, International Conference on Computer Science and Software Engineering, 2008.

[4] Zheng R, Li J, Chen H, Huang Z., “A framework for authorship identification of online messages: writing-style features and classification techniques”. Journal of the American Society for Information Science and Technology, February ;57(3), pp.378– 93, 2006.

[5] Zheng R, Qin Y, Huang Z, Chen H., “Authorship analysis in cybercrime investigation”, *In: Proc. 1st NSF/NIJ symposium. ISI Springer-Verlag*; pp. 59–73, 2003.

[6] de Vel O, Anderson A, Corney M, Mohay G., “Mining e-mail content for author identification forensics”, SIGMOD Record December ;30(4):55–64, 2001.

[7] Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, Mourad Debbabi., “Mining writeprints from anonymous e-mails for forensic investigation”, Digital Investigation, 2010.

[8] Stolfo S.J., Hershkop S., Ke Wang, Nimeskern O., “EMT/MET: systems for modeling and detecting errant e-mail”, *Proceedings of DARPA Information Survivability Conference and Exposition*, 2003.

[9] Stolfo S.J., Hershkop S., Ke Wang, Nimeskern O., Chia-Wei Hu, “Behavior-Based Modeling and Its Application to E-mail Analysis”,ACM Transactions on Internet Technology, Vol. 6,No. 2, May, Pages 187–221, 2006.

[10] Xiaoyan Fu_,Seok-Hee Hong,Nikola S. Nikolov,Xiaobin Shen,Yingxin Wu,Kai Xuk, “Visualization and Analysis of E-mail Networks”, Asia-Pacific Symposium on Visualisation, 2007.

[11] Fanlin Meng, Shunxiang Wu, Junbin Yang, Genzhen Yu, “Research of an E-mail Forensic and Analysis System Based on Visualization”, Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications, 2009.

[12] Sudhir Aggarwal,Jasbinder Bali,Zhenhai Duan,Leo Kermes,Wayne Liu,Shahank Sahai,Zhenghui Zhu, "The Design and Development of an Undercover Multipurpose Anti-Spoofing Kit (UnMask)", 23rd Annual Computer Security Applications Conference, 2007.

[13] Rabeah Al-Zaidy, Benjamin C. M. Fung, Amr M. Youssef, "Towards discovering criminal communities from textual data", *Proceedings of the 2011 ACM Symposium on Applied Computing*, 2011.

[14] Olivier de Vel, "Mining E-mail Authorship", KDD-2000 Workshop on Text Mining, August 20, Boston, 2000.

[15] S S.Appavu alias Balamurugan, Dr.R.Rajaram, "Data mining techniques for suspicious e-mail detection: A comparative study", IADIS European Conference Data Mining, 2007.

[16] D.V. Chandra Shekar and S.Sagar Imambi, "Classifying and Identifying of Threats in E-mails – Using Data Mining Techniques", *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol.I, IMECS, 19-21 March 2008, Hong Kong.

[17] Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas, "E-mail Mining: Emerging Techniques for E-mail Management",Aristotle University of Thessaloniki, Department of Informatics, Greece.

[18] Abbasi A, Chen H., "Writeprints: a stylometric approach to identity level identification and similarity detection in cyberspace", ACM Transactions on Information Systems, Vol.26, No.2, Article 7, March 2008.

[19] Jiexun Li, Rong Zheng, Hsinchun Chen, "From Fingerprint to Writeprint", Communications of the ACM, 2006.

[20] Sergio Decherchi, Simone Tacconi, Judith Redi, Fabio Sangiacomo, Alessio Leoncini and Rodolfo Zunino, "Text Clustering for Digital Forensics Analysis", Journal of Information Assurance and Security 5 (2010),pp.384-391.

[21] Gary Palmer, "A Road Map for Digital Forensic Research, "DFRWS Technical Report",Available:http://www.dfrws.org/2001/dfrwsrmfinal.pdf, 2001.

[22] Ryan Rowe, German Creamer, Shlomo Hershkop and Salvatore J Stolfo, "Automated Social Hierarchy Detection through E-mail Network Analysis",Joint 9th WEBKDD and 1st SNAKDD Workshop '07 August 12, 2007, San Jose, California, USA.

[23] M. Goldberg, M. Hayvanovych, A. Hoonlor, S. Kelley, M. Ismail, K. Mertsalov, B. Szymanski and W. Wallace, "Discovery, Analysis and Monitoring of Hidden Social Networks and Their Evolution", Technologies for Homeland Security, IEEE Conference, pp.1-6, 2008.

[24] Sobiya R. Khan, Smita M. Nirkhi, R. V. Dharaskar, "E-mail Mining for Cyber Crime Investigation", *Proceedings of International Conference on Advances in Computer and Communication Technology*, pp.138-141, February 2012.