International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
2nd National Conference on Innovative Paradigms in Engineering & Technology (NCIPET 2013) – www.ijais.org

E-mail Data Analysis for Application to Cyber Forensic Investigation using Data Mining

Sobiya R. Khan
P.G. Dept. of Computer Science & Engineering
GHRCE, Nagpur, (M.S), India

Smita M. Nirkhi
P.G. Dept. of Computer Science & Engineering
GHRCE, Nagpur, (M.S), India

R. V. Dharaskar, Ph.D
M. P. G. I,
Nanded,
(M.S), India

ABSTRACT
This paper discusses briefly the significance of e-mail communication in today's world, how substantial e-mails are with respect to obtaining digital evidence. The framework proposed by authors employs state-of-the-art existing data mining techniques. Experiments are conducted for e-mail analysis on the Enron data corpus. The intent of the proposed system is to provide assistance during forensic investigation. In this paper, we enhance the results obtained in our previous work on statistical analysis and provide our findings on e-mail classification experiments.

General Terms
Data Mining, Machine Learning, Digital Forensic.

Keywords
E-mail forensic analysis, Statistical Analysis, Classification and Clustering techniques, Authorship identification, Social Network Analysis.

1. INTRODUCTION
In this Internet age, e-mail still holds a significant place at the summit, in finding useful digital evidence in mountains of data. It is like an inseparable accessory rather than a tool for most of us, with trillion e-mails sent a year, making it a key evidential element in almost every case litigated today. We are more prone to using e-mail rather than talking on telephones or putting things to paper. So be it a software firm, a bank, or an institution, e-mails are everywhere! However, today new digital trends like social networking, texting and tweeting are slowly nibbling e-mail as our primary communication, but still they are no less digital. So every time when our keyboard is doing the talking, we are naively leaving digital traces behind. Forensic experts are masters of finding and following these traces, rightfully to the evidence. Civil and criminal court proceedings are seeking reliable digital evidence to punish the convict in cyber crime cases. For this reason, expert forensic analysis of e-mails and other electronically stored information is paramount when evidence goes digital. This had lead to the need for efficient automated tools in the hands of forensic experts [29].

1.1 Identification and Extraction
The first step in any e-mail analysis is identification of e-mail sources and how the e-mail servers and clients are used in an organization. E-mail clients and servers have expanded into full databases, document repositories, contact managers, time mangers, calendars' and many other applications, more than just a way of sending messages. Organizations use these powerful, database enabled e-mail and messaging servers to manage cases, track clients and share data. Most users are still unaware that even though they have clicked the delete button or emptied their Deleted Items folder, many times e-mails can be forensically extracted even after deletion. It's common policy for most organizations, such as banks or brokerage firms, to have e-mail archiving for regulatory purposes, with users being totally unaware, that all their e-mails are being stored for years in a searchable, retrievable format. Most users are oblivious to the fact that e-mails may reside on servers unbeknown to them, or on backup tapes that were created during the normal course of business. Certainly, they may also be extracted from the hard disk of the client or the server as well. If properly conducted and managed the forensic analysis of e-mail yields documents that can be easily correlated by date, subject, recipient or sender and yield a highly understandable and easy to follow map of events and entities. This could play an evidentiary role in criminal cases pertaining to cyber crimes. However manual analysis of this enormous data is impractical. In this context, Data mining and machine learning techniques have reliably paid off. The framework in [29-31] proposed by authors is based on these well established techniques and efforts to provide a better insight in e-mail analysis by assisting the forensic investigator during initial stage of any forensic investigation.

In this paper, we extend the implementation of our proposed framework in [29-31]. This paper is divided into four sections. Section II briefly summarizes the related work in e-mail mining. Section III extends the statistical findings obtained in [30] and the experimental results on e-mail classification. Section IV presents the conclusion drawn.

2. RELATED WORK
E-mail mining is an upcoming research area wherein researchers are constantly striving to find upbeat methods to restrain against the increasing e-mail abuse. Digital Forensic technology has already become a centre of attention among researcher's and law professionals to help curb the amount of cyber crime [28]. Existing techniques of data mining, machine learning algorithms and visualization have been used profusely by researchers in their endeavor. Researchers have brought up various tools and frameworks for e-mail analysis from time to time, to aid the forensic investigator, which are outlined in [1], [18-22]. Data mining techniques of classification and clustering have given credible results in e-mail forensic analysis as explored in [1-3], [23]. Social Network analysis of e-mails helps in determining user behaviors and their social circle. Valued forensic analysis has been obtained on e-mail social network analysis in [1], [24-27]. E-mail authorship analysis is an emerging research avenue with respect to forensic investigation. Authorship analysis, draws its roots from linguistic research, and has gained momentum with respect to e-mail and online message authorship over the time course and commendable results

have been obtained by researchers as discussed in [4-15]. Another aspect added with respect to linguistic research of e-mail is gender identification of e-mails which could be of vital forensic value, and is discussed in [16-17]. However, presently the authors haven't included gender identification feature in their proposed framework due to time constraints. Extensive details of the related literature on the proposed framework can be followed in [29, 31], where the authors have thoroughly explored the related work on e-mail analysis.

3. EXPERIMENTAL RESULTS

The proposed framework employs data mining techniques to achieve the myriad functionalities. The framework is proposed to perform Statistical Analysis, Classification & Clustering, Author Identification and Social Network Analysis. The intent of the work is to overcome the confines observed in previous systems. To evaluate our implementation, we are using the Enron e-mail corpus made available by MIT at http://www.cs.cmu.edu/~enron/. The proposed framework is implemented in Java and uses the data mining tool weka as previously discussed in [29-30].

3.1 E-mail Statistical Analysis

Statistical analysis of e-mail data is calculated from e-mail ensembles, and the communication pattern so obtained reflects a great deal of information valued 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. [29, 30]. In our previous work in [30] we had calculated limited statistics. Here we have extended the statistics calculated per individual e-mail user. Also the results have been enhanced using graphical bar charts. For case study we are viewing the statistical analysis of user Allen P. Figure 1-5 show the various statistics calculated over the inbox and outbox of the user Allen P. The inbox and outbox statistics reveal the persons with whom Allen P is frequently communicating. We can also infer the least communicating entities from the same graph. Thus the outliers can be further analyzed by the forensic investigator for detailed analysis.

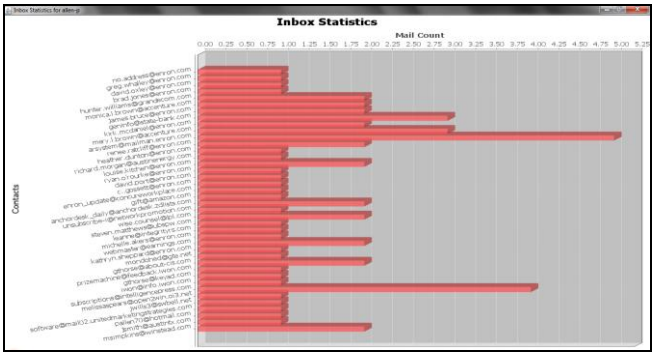


Fig 1: Inbox Statistics

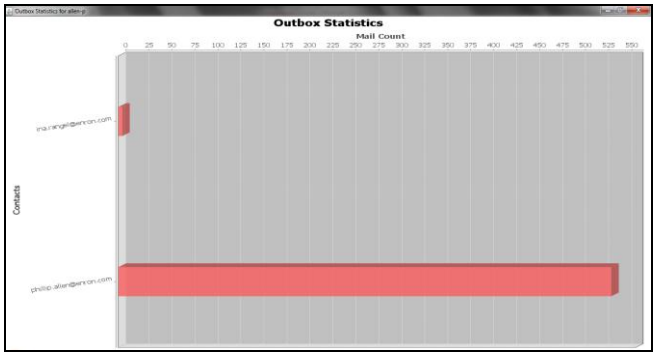


Fig 2: Outbox Statistics

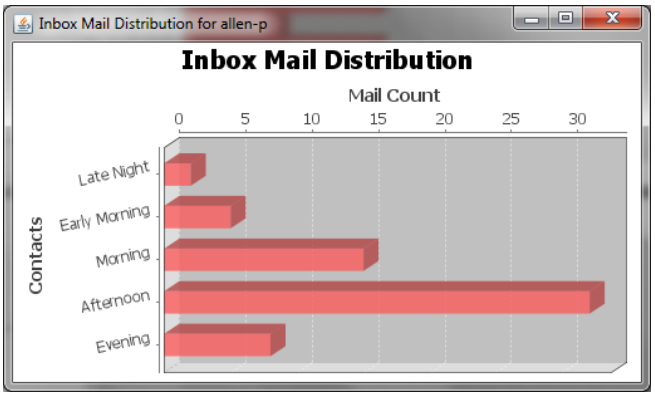


Fig 3: Inbox Mail Distribution

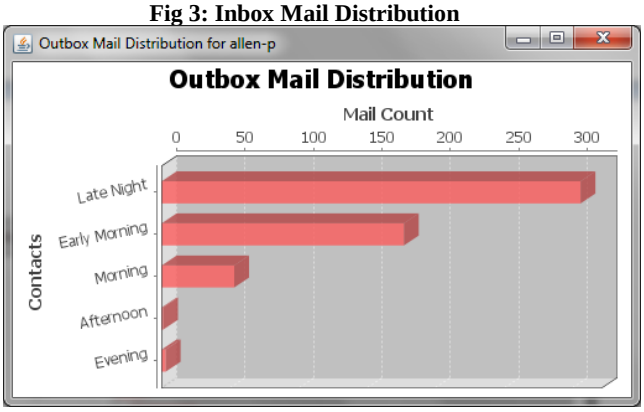
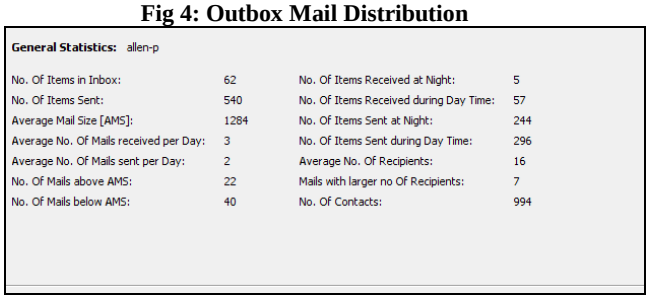


Fig 4: Outbox Mail Distribution



General Statistics: allen-p

No. Of Items in Inbox:	62	No. Of Items Received at Night:	5
No. Of Items Sent:	540	No. Of Items Received during Day Time:	57
Average Mail Size [AMS]:	1284	No. Of Items Sent at Night:	244
Average No. Of Mails received per Day:	3	No. Of Items Sent during Day Time:	296
Average No. Of Mails sent per Day:	2	Average No. Of Recipients:	16
No. Of Mails above AMS:	22	Mails with larger no Of Recipients:	7
No. Of Mails below AMS:	40	No. Of Contacts:	994

Fig 5 : General Statistics

Similarly, the inbox and outbox distribution at over various hour of the day provides a summary about the communication pattern of a certain user. As identified in this case, Allen P is frequently sending mails at late hours in night. Such information can be of crucial use to the forensic investigator in cases of threat or fraud from a disgruntled employee.

Rest other statistics can also be shown graphically but have been skipped due to space issues. Thus, statistical analysis can provide an initial insight to the forensic investigator at the first

glimpse of the e-mail ensemble. This can result in confining the number of suspects or even identification of possible suspects at the primary stage of the investigation.

3.2 E-mail Classification

Most e-mail mining tasks are being accomplished by using e-mail classification at some point. E-mail classification is the assignment of an e-mail message to one of the category, from a pre-defined set of categories. Examples of applications are automatic mail categorization into folders; spam filtering and author identification. Generally classification is performed in two steps: data cleaning which is followed by features extraction. Next the extracted features are bifurcated to represent the training and test sets. The training sets are given class labels and it is used to develop a classifier model. The accuracy of the classifier depends on the aptness of the training data. The developed model is then tested with the test data and the classifiers accuracy is tested. Sometimes some other data can also be used in order to validate the developed model.

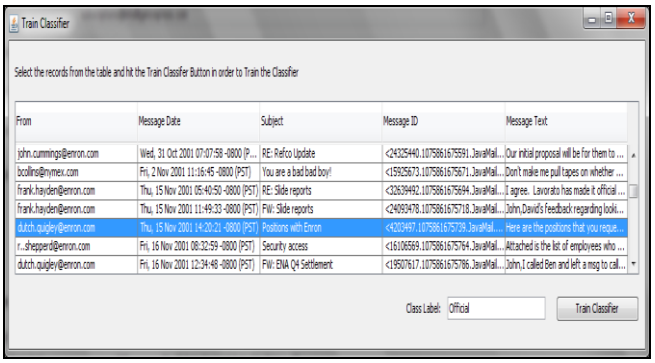


Fig 6: Training of Classifier

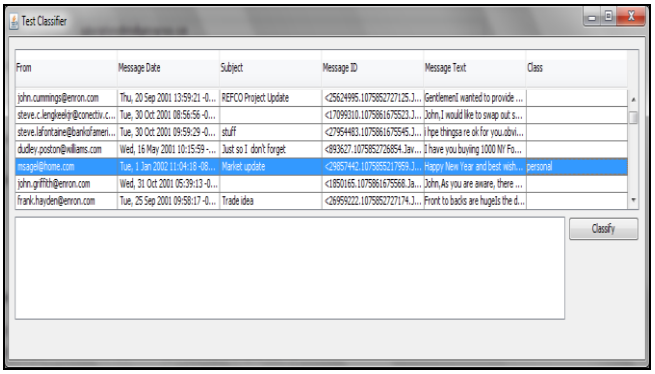


Fig 7: Testing of Classifier

Here we will be classifying the e-mails based on their topic into two classes: those dealing with company business (official) and those that were personal specifically for Enron dataset. For Classification, we are using Decision Tree classifier represented as J48 in the weka classes, which has shown promising results in most of the studies. For e-mail classification, the body of e-mail is converted to a vector of metrics called features. Using java *StringToWordVector* API of weka, each character stream is converted into distinct tokens or words. Some of the words may appear in different forms (for instance, verb, noun, and adjective, etc.), or different tenses such as present, past, and future). Such words are stemmed to their common root. For example 'report', 'reported', and 'reporting' will be stemmed to their root form 'report'. We are using weka to internally stem the words to their base form. For feature selection we are using the *tf-idf* measure and weka's internal attribute selection methods in order to reduce dimensionality of the feature set.

For experiment we considered a subset containing around 200 e-mail messages classified manually into two classes: those dealing with company business (official) and those that were personal. Each category contained 100 e-mails. We randomly created the training set and developed the classifier model using J48 Decision tree class of weka. We constructed a training set by randomly selecting 60 e-mails from each class, while the remaining 40 e-mails were used for testing. Figure 7 shows the testing procedure of our framework. This same experiment was repeated for several random training set in order to test the robustness of our classifier. The precision of the classifier varied between 74% and 89%, with an average precision of about 82%. The classifier's precision is computed to measure the accuracy of the developed model, and is calculated as the percentage of true positives. The observed results from our developed classifier model are within acceptable range and are showing promising result.

4. CONCLUSION

With growing e-mail abuses investigators require efficient automated tools for fast e-mail analysis to help the forensic investigators gather evidence. Automation can ease the process, help in saving the time and can help to nab the culprit in time. E-mail has become vital for inter-personal communication and professional life, so efficient solution for fast e-mail analysis is the need of the hour. Through statistical analysis we have concluded that it could be effective at the primary stage of forensic investigation in order to identify prime suspects or limit the number of suspects by identifying anomalous behavior. However in order for this to be full proof, the context of the data should be kept in mind which is very necessary. The experiments on e-mail classification have shown promising results. The results can be compared with other classifiers but have been limited due to time constraints. The authors propose to incorporate this as a part of future scope. The implementation of the rest of the proposed framework is ongoing.

5. 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.

[11] Gray, A., Sallis, P., & MacDonell, S., "Software forensics: Extending authorship analysis techniques to computer programs", Third biannual conference of the International Association of Forensic Linguists (IAFL '97), 1997.

[12] Argamon, S., S ˇ ari´c, M., & Stein, S.S., "Style mining of electronic messages for multiple authorship discrimination", Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 475–480). ACM Press, 2003.

[13] Chaski, C., "Empirical evaluations of language-based author identication techniques",Forensic Linguistics, 8, 2001.

[14] Gui-Fa Teng'J, Mao-Sheng Lai I, Jian-Bin Ma', Ying Li, "E-mail Authorship Mining based on SVM for Computer Forensic", Proceedings of the Third International Conference on Machine Learning and Cybermetics, Shanghai, August, pp.26-29, 2004.

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

[16] Corney, M., de Vel, O., Anderson, A., & Mohay, G. , "Gender-preferential text mining of E-mail discourse", Eighteeth annual Computer Security Applications Conference (ACSAC 2002), Las Vegas, NV, 2002.

[17] Koppel, M., Argamon, S., & Shimoni, A.R., "Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401–412, 2002.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] Sobiya R. Khan, Smita M. Nirkhi, R. V. Dharaskar, "Mining E-mail Content for Cyber Forensic Investigation", UACEE International Journal of Computer Science and its Applications, Vol. 2, Issue-2, pp.112-116, Aug 2012.

[31] Sobiya R Khan, Smita M Nirkhi and R V Dharaskar, "Author Identification for E-mail Forensic", IJCA Proceedings on National Conference on Recent Trends in Computing NCRTC(2):29-32, May 2012.