A Methodology for Cyber Crime Identification using Email Corpus based on Gaussian Mixture Model

V.Sreenivasulu
Department of Computer Science and Engineering
Gandhiji Institute of Science and Technology,
Krishna District - Andhra Pradesh - India

R. Satya Prasad, PhD
Department of Computer Science and Engineering
Acharya Nagarjuna University - Guntur,
Andhra Pradesh - India

ABSTRACT

The area of crime investigation has extended its roots to cyber media and has emerged exponentially with the technological strides. Among the various media used in Digital Forensics, Email Forensics took up the leading segment. In order to investigate the cyber crimes, there is an immense need to analyze the bulky email gatherings forensically. Data mining methods help in analyzing these large collections of data. Mixtures of data mining models along with the related methodologies are proposed in this paper to facilitate the email forensic assessor. The Performance is evaluated using False Rejection Ratio (FRR) and False Acceptance Ratio (FAR).

Keywords

Email Forensics, Data Mining, Digital Forensics, Word Net, Query Analysis.

1. INTRODUCTION

The rapid expansion of science and technologies has helped the mankind in witnessing new methods for leading a better quality of life. At the same time these technological updates have paved new approaches for the criminals and anti-social elements in indulging the criminal activities using the cyber media as a medium. To prevent these activities, effective methods are needed. Among the various methods of identifying the cyber criminals, analysis of the emails is considered to be more effective. The practice of emails for the fraudulent actions is also emerging with an elevated speed. Forensic Analysis of these emails can probe or prove a crime committed. Email forensics can appropriately be defined as the usage of focused techniques used for the collection, protection and analysis of emails with a vision to submit the evidences in a court of law. As the usage of internet technologies are expanding, the usage of emails has been increased and investigating these bulk emails is a great task for the forensic examiners. Therefore new methods are to be evolved for the effective analysis of these emails to uproot the inherent information concealed in the emails.

The core idea behind the digital forensics procedures is to identify the relevant information from the digital media and which can be used as evidence in proving a crime. To recognize and interpret the meaningful information hidden in the emails, Information Retrieval (IR) plays a vital role. Information Retrieval systems recognize the relevant information by retrieving the data based on the user’s queries. In this paper we recommend a system to retrieve the most related emails from the bulk email collections of emails and exhibit these emails to the forensic examiners for better understanding. Every piece of information is of core essence in the investigation process, so it is needed to convert the information retrieved into a meaningful format so that it will be the utmost usage for the Forensic examiners. Every forensic information retrieval systems work with the aim towards high recall ratio. The present system addresses the issue of high recall using the concepts of interpreting the semantic meaning of the words within the emails for which in the present paper the concepts of ontology i.e. Word Net is used.

In order to have an effective examination, the forensic examiners should identify the emails which have a high degree of matching with the relevant information; this is assumed to be the primary goal of any system which aims at High recall. In this paper, we present a novel methodology, where the emails from a particular user under examination are scanned and the text from each of these emails is extracted and the contents are indexed. This indexing list is given as input to the semantic analyzer using Word Net and the related synonyms of the words are listed. This list is maintained in a book called code book. Further, for each of the words, the grammar is also extracted; the relevant acronyms are expanded using the word Net. Concepts of clustering summarization are used. The rest of the article is organized as follows, in Section-2; the related work in this area is highlighted, in Section-3; the dataset considered is presented along with a sample from the corpus. The methodology is highlighted in Section-4 of the paper; and Section 5; concludes the paper.

2. RELATED WORK

The restriction of the message sharing and content sharing is a concern because of the deceptive activities taking place all around the globe. Therefore, it is needed to counter attack these fraudulent activities to curtail the criminal activities. Lot of research has been reported in the literature regarding methodologies proposed to curb the fake emails and disrupting the activities. Data mining techniques of classification and clustering have given unbelievable results in email forensic analysis as explained in [1][2][3]. Among the available filters, Naïve Bayesian is a basic Standard Statistical approach on probability proposed by Sahami et al [4]. The algorithm assumes the classification of new Email by identifying it as spam or legitimate. It achieves the features using a training set that has already been pre classified correctly and then checks for specific work that appears in the email. The indication of high probability says that the new one is a spam email. In recent years the authorship identification has been used in diverse number of application areas some examples of identifying authors in written literature, in the program source code [5] and in forensic analysis for criminal cases. Some techniques involve identification of Spoofed E-Mails [6] based on various techniques of SPF (Sender Policy Framework), Sender ID, Domain Keys Identified Mail (DKIM). In most of the approaches, tools were used to surmount the issue. Some of the significant contributions made by eminent researchers are presented in this section. The researchers have developed a tool helpful in analyzing the email content and also helping out mechanisms to reveal the related information necessary for the investigative procedures during the forensic investigations. The main advantage of the proposed tool is that, it overcomes the disadvantages of the ongoing forensic search mechanisms, where the search is confined by grouping or inappropriate filtering. The authors have also proposed a new methodology for indexing the contents in the emails, where by the time complexity for searching the related documents from a particular email minimizes. The [7] investigators have proposed a methodology, for retrieving the contents from the emails, in which ranks were assigned to each of the words used in the emails, and the ranking mechanism is claimed to be novel. Once the users present small queries, in turn the retrieval procedures underlined could not be able to display the relevant information. Some have preferred towards the usage of WorldNet tool to address this issue. [8] Have utilized the Enron email dataset for the experimentation purpose. In [9], authors showcased a methodology based on a tool developed on data mining to envisage a variety of analyses from the email. The retrieval of relevant location and address of the users based on IP address extracted from the email address is discussed by the authors in [10]. The authors in their paper [11] tried to analyze the emails based on the SMTP protocols and HTTP protocols. In [12], authors have proposed techniques for discovering emails in one conversation, capturing the conversation structure and summarizing the email conversation. The analysis of the emails based on the phishing attacks is presented by [13] using the UNIX tools system. The methodology of identifying the users pattern based on clustering and classification techniques is addressed by the authors in [14]. In this article, we propose a methodology for addressing the email forensics using the data mining techniques.

3. DATA SET

To present the model the emails, Enron data (https://www.cs.cmu.edu/~./**enron**/) which contains a repository of emails set is considered. This data set contains about 5, 17,431 emails. These emails are considered for testing the fraud detection.

3.1 Sample Email Corpus from ENRON Data Set

The sample email obtained from the Enron Data set is presented below

A couple of the entities have shown definitive interest in hiring the group, but our Enron and employment contracts have created significant issues.

A meeting is a good idea.

A note on the front door, “the Body Shop would be shutting down at 7:30 p.m. Thursday..”.

A quick question before our meeting…

A substantial advertising discount is being made available to the corporation that sponsors this documentary.

After all the above, I just have to say...we really do appreciate your help resolving all these issues...

After such 10 day period, the parties shall be obligated to resume performance as if the event of force majeure had been remedied, and if such party does not resume performance, such nonperformance shall be deemed a Triggering Event allowing the other party to terminate.

After sunning and drinking all day we feasted on a fleisch dinner...that was all we had here to we grilled bauchfleisch, steaks and bratw?rst..., good greasy stuff.

After the session, I had a discussion with you.

After you set your curves(do not hit finished file), go to the GD Pmgr tab and change the date in the left hand corner to two business days out(Friday set for Tuesday).

All of the others are charging 3% fees for matching buyers and sellers in a bulletin board or auction

4. METHODOLOGY

The proposed system architecture of the model is presented in the following

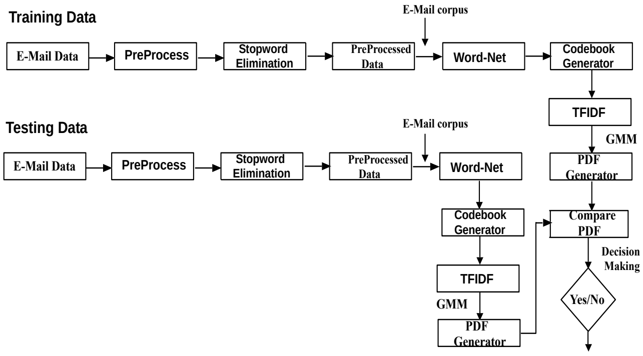
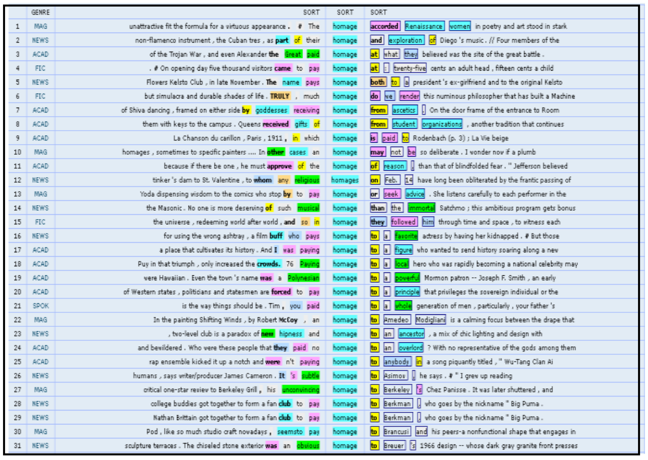


Figure-1: Proposed System Architecture

The emails extracted from the ENRON dataset are analyzed and each mail was processed using the Data mining concepts like tokenization, preprocessing where the stop words are eliminated and the indexing is performed. Each of the words may have different synonyms, so the Word Net is used to extract the meaning of each of the words and now, each word is indexed and a rank is assigned along with a group of words. A code book is used for this purpose. The acronyms are also taken care by utilizing the Ontological semantics. When a user is assumed to be a fraud, the email forensics is carried out and the query mail is compared to each of the emails in the indexed list. While checking for the authentication, the indexing is considered. These emails will be presented to the user as well as forwarded to the further modules for in-depth analysis**.**

Processed data using the Word Net



	GENRE	SORT	SORT	SORT
1	MAG	unattractive fit the formula for a virtuous appearance . # The	homage	accorded Renaissance women in poetry and art stood in stark
2	NEWS	non-flamenco instrument , the Cuban tres , as part of their	homage	and exploration of Diego 's music . // Four members of the
3	ACAD	of the Trojan War , and even Alexander the Great paid	homage	at what they believed was the site of the great battle .
4	FIC	. # On opening day five thousand visitors came to pay	homage	at : twenty-five cents an adult head , fifteen cents a child
5	NEWS	Flowers Kelsto Club , in late November . The name pays	homage	both to a president 's ex-girlfriend and to the original Kelsto
6	FIC	but simulacra and durable shades of life . TRULY , much	homage	do we render this numinous philosopher that has built a Machine
7	ACAD	of Shiva dancing , framed on either side by goddesses receiving	homage	from ascetics . On the door frame of the entrance to Room
8	ACAD	them with keys to the campus . Queens received gifts of	homage	from student organizations , another tradition that continues
9	ACAD	La Chanson du carillon , Paris , 1911 , in which	homage	is paid to Rodenbach (p. 3) ; La Vie beige
10	MAG	homages , sometimes to specific painters .... In other cases an	homage	may not be so deliberate . I wonder now if a plumb
11	ACAD	because if there be one , he must approve of the	homage	of reason , than that of blindfolded fear . " Jefferson believed
12	NEWS	tinker 's dam to St. Valentine , to whom any religious	homages	on Feb. 14 have long been obliterated by the frantic passing of
13	MAG	Yoda dispensing wisdom to the comics who stop by to pay	homage	or seek advice . She listens carefully to each performer in the
14	NEWS	the Masonic . No one is more deserving of such musical	homage	than the immortal Satchmo ; this ambitious program gets bonus
15	FIC	the universe , redeeming world after world , and so in	homage	they followed him through time and space , to witness each
16	NEWS	for using the wrong ashtray , a film buff who pays	homage	to a favorite actress by having her kidnapped . # But those
17	ACAD	a place that cultivates its history . And I was paying	homage	to a figure who wanted to send history soaring along a new
18	ACAD	Puy in that triumph , only increased the crowds. 76 Paying	homage	to a local hero who was rapidly becoming a national celebrity may
19	ACAD	were Hawaiian . Even the town 's name was a Polynesian	homage	to a powerful Mormon patron -- Joseph F. Smith , an early
20	ACAD	of Western states , politicians and statesmen are forced to pay	homage	to a principle that privileges the sovereign individual or the
21	SPOK	is the way things should be . Tim , you paid	homage	to a whole generation of men , particularly , your father 's
22	MAG	In the painting Shifting Winds , by Robert McCoy , an	homage	to Amedeo Modigliani is a calming focus between the drape that
23	NEWS	, two-level club is a paradox of new hipness and	homage	to an ancestor , a mix of chic lighting and design with
24	ACAD	and bewildered . Who were these people that they paid no	homage	to an overlord ? With no representative of the gods among them
25	ACAD	rap ensemble kicked it up a notch and were n't paying	homage	to anybody in a song piquantly titled , " Wu-Tang Clan Ai
26	NEWS	humans , says writer/producer James Cameron . It 's subtle	homage	to Asimov . he says . # " I grew up reading
27	MAG	critical one-star review to Berkeley Grill , his unconvincing	homage	to Berkeley 's Chez Panisse . It was later shuttered , and
28	NEWS	college buddies got together to form a fan club to pay	homage	to Berkman , who goes by the nickname " Big Puma .
29	NEWS	Nathan Brittain got together to form a fan club to pay	homage	to Berkman , who goes by the nickname " Big Puma .
30	MAG	Pod , like so much studio craft nowadays , seemsto pay	homage	to Brancusi and his peers-a nonfunctional shape that engages in
31	NEWS	sculpture terraces . The chiseled stone exterior was an obvious	homage	to Breuer 's 1966 design -- whose dark gray granite front presses

4.1 Algorithm

The proposed architecture of the model is presented as follows.

Step 1:

The email datasets are extracted from the ENRON dataset and preprocessed to eliminate the missing data and eliminate the stop words. The preprocessed data is given as an input to the semantic tool, WORDNET. The purpose of this tool is to analyze the grammar, identify the term index, acronyms and this data is also used for identifying the different synonyms for a given keyword. The synonyms in this paper are considered only for the adjectives. A code book is generated for the group of adjectives and the indexing is carried out for each word in the code book. This indexed data for email is given as input to the Gaussian Mixture Model (gmm). The main advantage of considering gmm is that the patterns inside the data can be clearly estimated. It is symmetric model and can be very useful for analyzing huge data sets. This mixture model is very effective for modeling the data having very low index and very high index. The probability function of the Gaussian Mixture Model is given by

$$\mathcal{N}(X|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

... (1)

Step 2:

The indexed data is given as input to the gmm presented in equation (1) and the corresponding probability density function (pdf) is estimated. The process is repeated for all the data in the email and all the emails in the dataset. This is called training data. For testing purpose the same procedure is adopted and pdf of the query email is generated. The pdf is compared with the pdfs in the training data using KL divergence (Khaul Loullie Formulae). It is used to identify the relation between the email in the database and the email which happens to be false email i.e., the suspected email. MathCAD is used for the calculation purpose where the log of A and B are first calculated and then the integration is carried out. The MathCAD is used to perform the mathematical calculation of integration, differentiation and volume integrals. The alternative to K L Divergence is identifying the likelihood estimate. The formulae for calculating the KL divergence is given by

$$KL\ (A,B) = \int A(x)\ log\left(\frac{A(x)}{B(x)}\right)dx$$

... (2)

Where A(x) is the corpus of a person x and B(x) is the email which is under consideration.

Step 3:

For each query email, the relevant PDF is retrieved. The process is repeated for all the query emails to find the best fit. The results derived are presented below.

Step 4:

To evaluate the performance of the given model, we used metrics like False Acceptance Rate (FAR) and false Rejection Rate (FRR).

Table 1 The accuracy, FAR and FRR of Emails

Trait	Algorithm	Acceptance Rate (%)	FAR (%)	FRR (%)
Email-set1	GMM	88.70	0.07	0.93
Email-Set2	GMM	91.05	0.064	0.936

The above outputs evaluated are compared using Acceptance Rate and False Acceptance Rate. The FAR is calculated using

$$\frac{[(Dataset\ under\ consideration)-(Total\ Number\ of\ related\ words]}{Size\ of\ total\ data\ set}\ X\ 100$$

The acceptance rate is calculated as

$$\left[\frac{Total\ Number\ of\ items}{Number\ of\ items}X\ 100\right]$$

5. CONCLUSION

In this paper a novel methodology for email forensics is highlighted using the concepts of data mining, semantic ontologies and Gaussian mixture model. The outputs derived are tested against accuracy using metrics like FAR and FRR. The results derived are presented above and revealed that the proposed methodology possesses high false acceptance rate and low false rejection rates. This methodology can be very useful in identifying from email corpus and thereby helping to identify the law breaker.

6. REFERENCES

[1] Rachid Hadjidj *et al* “Towards An Integrated Email Forensic Analysis Framework”, Digital Investigation 5,pp.124-137, 2009.

[2] S.Appavu alias Balamurugam, Dr.R.Rajaram,”Data mining techniques for suspicious email detection: A comparative study”,IADIS European Conference Data Mining, 2007.

[3] D.v. Chandra Sekhar and S. Sagar Imambi,” Classifying and Identifying of Threats in Email Using Data Mining Techniques”, Proceedings of the International MlitiConference of Engineers and Computer Scientists Vol.1,I IMECS,19-21 March 2008,Homg Kong.

[4] Sahami *et al* “A Bayesian Approach to Filtering Junk Email In Learning for Text Categorization” – papers from the AAAI Workshop, pp. 55-62, Madison Wisconsin AAAI Technical Report WS-98-05, 1998.

[5] Gray *et al*,” Software forensics: Extending authorship analysis techniques to computer programs”, Third biannual conference of the international Association of Forensic Linguists (IAFL’97),1997.

[6] Dhanalakshmi R,L.Kavisankar,C.Chellappan“Enhanced Email Authentication Against spoofing Attacks To Mitigate Phishing, European Journal of Scientific research Vol . Issue 2011

[7] Marwan Al-Zarouni. (2004). “Tracing E-mail Headers”, Proceedings of Australian Computer, Network & Information Forensics Conference on 25th November, School of Computer and Information Science, Edith Cowan University Western Australia 2004, pp. 16-30.

[8] eMailTrackerPro, http://www.emailtrackerpro.com/

[9] EmailTracer, http://www.cyberforensics.in

[10] Adcomplain, http://www.rdrop.com/users/billmc /adcomplain.html

[11] Aid4Mail Forensic, http://www.aid4mail.com/

[12] AbusePipe, http://www.datamystic.com/abusepipe .html

[13] AccessData's FTK, http://www.accessdata.com/

[14] EnCase Forensic, http://www.guidancesoftware.com

[15] FINALeMAIL, http://finaldata2.com

[16] Sawmill-GroupWise, http://www.sawmill.net