Swarm Intelligent Agents of E-mail Classification for Contribution of Cyber and Forensic System

Assistant Prof. Dr. Abdulkareeem Merhej Radhi
Al-Nahrain University, Information Engineering College, Baghdad-Iraq
Communication and Information Engineering Department

Abstract

Due to the Internet is an open, direct transmission tool and available to all for transmitting and exchanging messages, including those circulating among criminals, terrorists and those with improper motives towards others, therefore it became necessary to classify these messages by intelligent agent as a forensic evidence. In light of this, this paper included the urgent need to find a self-classified analysis of e-mail and to clarify what is suspicious from it, then classify the fraud Issuers which is covered in this paper. Sample data is a set of real e-mails. Some virtual messages were added to enhance the accuracy of the results obtained. The proposed work relies on swarm intelligent agents and modification of Voronoi algorithm such that the issues of the messages, including suspicious messages, are divided into communities. Moreover, Communities are divided into categories, each given a specific rank, depending on the quality and size of the threat messages. Moreover, this technique can be applied to messages and comments from people who talk about a particular subject and participants in the social media pages.

Keywords: Forensic; Voronoi; Classifier; SPF, Entropy.

1.INTRODUCTION

Today we are living in the information age; all the information which is transferred over the internet is through the digital devices. With the advent of the World Wide Web, advanced forms of digital crimes came into picture. Criminal uses the Digital devices to commit Digital crime, so for the investigation forensic Experts have to adopt practical frameworks and methods to recover data for analysis which can comprise as evidence. Investigation of Digital forensics adopts three essential processes: Data Generation, Data Preparation and Data warehousing. Data Mining has unlimited potential in the field of Digital Forensics. Computer forensics is an emerging discipline investigating the computer crime [1]. The Internet provides a convenient platform for cyber criminals to anonymously conduct their illegitimate activities, such as phishing and spamming. As a result, in recent years, authorship analysis of anonymous emails has received some attention in the cyber forensic and data mining communities [2]. The goal of Digital forensics process is to preserve any evidence in its most original form while performing a structured investigation by collecting, identifying and validating the digital information for the investigation of particular digital crime. Email is the most common mode of communication today. Email not only used for sending messages/text only but also to send audio, video and other files as an attachment. It is the main resource for business communication. As it is most popular and common mode of communication on the internet, it also attracts criminals or persons having mischievous intent. Cyber criminals misuse it for sending spam, threats, Phishing-emails, propagating malicious software like virus and worms, distributing illegitimate material like child pornography, hoaxes and also used for other criminal activities. So it is necessary to secure our email system and also to identify criminal, collect evidence against them and punish them under the court of law [3].

2. PREVIOUS WORK

Iqbal and Khan [2]-2010, address the problem of authorship verification of textual documents and employ detection measures that are more suited in the context of forensic investigation; they borrow the NIST’s speaker recognition evaluation (SRE) framework. Khan, Nirkhi and Dharaskar[4]-2013 propose to employs state-of-the-art existing data mining techniques. Experiments are conducted for e-mail analysis of the Enron data corpus. The intent of the proposed system is to provide assistance during the forensic investigation. They enhance the results of statistical analysis. Kayarkar, Nirt, and motwani[1]-2014 introduce the Cyber Forensics using Sequence Mining algorithm, by comparing it with association rule mining algorithm parameters. Chhabra and Bajwa[3]-2016 review working and architecture of current email system and the security protocols followed generally to secure our email communications and the limitations they contained, further email forensics which is a process to analyze e-mail contents, header information, transit path for email, sender or receiver information and other details to collect evidence.

3. FORENSICS AND E-MAILS

Most countries recognize email as legitimate document evidence. Emails have been used as substantial sources of evidence in cases of homicide, cyber stalking, harassment, spoofed identity and espionage. The digital forensic aspect of emails (e-mail forensics) requires urgent attention, due to its impact in solving most of the cases of Computer Frauds and Cyber Crimes (CFCC). To make things worse, investigative and law enforcement agencies are underprepared to tackle the explosion of this new unseen, unheard, and innovative way of committing the crime. Technologies such as quantum computing, DNA computing, and “Adaptive or Reconfigurable Computing,” make hardware behave the flexibly and can be tailored to imitate various stipulations [6].

4. EMAIL ANALYSIS

E-mail may be considered as one of the main vectors of Internet communication, and its traffic has been increasing tremendously since the advent of the World Wide Web. Nowadays, it is estimated that more than two trillion E-mail messages are sent per year either for business or personal use. With the increase in E-mail traffic, the exploitation of E-mail for illegitimate practices is also becoming a common practice. Some of the known examples are: sending spam, threats, hoaxes, bullying, harassment, racial vilification. viruses and worms[5]. One of the major difficulties facing E-mail investigators is the large amount of E-mails to examine and the reliability of the information contained in these E-mails. E-mail, by its nature, is very easy to send. The "From" address header field can be easily forged as in ease of spam E-mails. The metadata containing the E-mail header and the path along which the message has traversed can also be forged or kept anonymous. An E-mail can also be routed through anonymous Email servers to hide any information about its origin. Sometimes the only useful information to help identify the author of an E-mail is the E-mail structure and the message it carries. Several existing commercial solutions analyze and partly “understand” the email content. For instance, Gmail focuses mainly on context sensitive advertisements, but can also detect events, addresses or package tracking numbers. Similarly, Zim-bra or Clear Context tries to recognize some objects in the text. Xobni focuses mainly on the extraction of contact data from email signatures [7].

5. E-MAIL CLASSIFICATION

To recognize illegal emails in data set of different emails, classification is the main process for satisfying the aim of receiving forensic evidence. There are different models to classify label E-mails. Hershkop[8], introduce models consist of traditional information retrieval and text classification models. Some common terms used in the text: Features, target function or target label (class label), false positive rate, False negative rate, sample error rate, true error rate, and bias. Different algorithms used in classifying E-mails. Byes, N-Gram,TF-IDF,User Clique,...etc.

5.1. Classification Methods

Different classification methods can be adapted to classify emails from the huge data set. They mainly differ in the statistical assumptions made of the data and the type of algorithms needed to construct the classifier [9].
They divided into three main categories: Linear classifier, Nearest Neighbor classifier, and Classification trees. They categorized to supervised and unsupervised learning algorithms. To classify data requires training unlabeled examples and tested according to weighted scoring function.

5.2. Feature Selection for Classification

The first phase of virtually all classification algorithms is that of feature selection. In most data mining scenarios, a wide variety of features are collected by individuals who are often not domain experts. Clearly, the irrelevant features may often result in poor modeling, since they are not well related to the class label. In fact, such features will typically worsen the classification accuracy because of over fitting, when the training data set is small and such features are allowed to be a part of the training model [10].
Michie, Spiegelhalter, and Taylor[11], introduce distinguish three common cases, only the first leading to what statisticians would term classification: Classes correspond to labels for different populations, Classes result from a prediction problem and Classes are pre-defined by a partition of the sample space.

6. PROPOSED ALGORITHM

This research introduces unsupervised classification procedure with modified Voronoi classification algorithm to classify data set target of E-mails to different classes. The Classified classes introduced for training phase and then tested.

6.1. Voronoi Classifier

A *Voronoi* diagram may be generated by any finite set of points in a plane. The partitioning of a plane with *n* points into *n* convex polygons such that each polygon contains exactly one point and every point in a given polygon is closer to its central point than to any other [12]. Figure [1] depict Voronoi diagram with labeled regions.

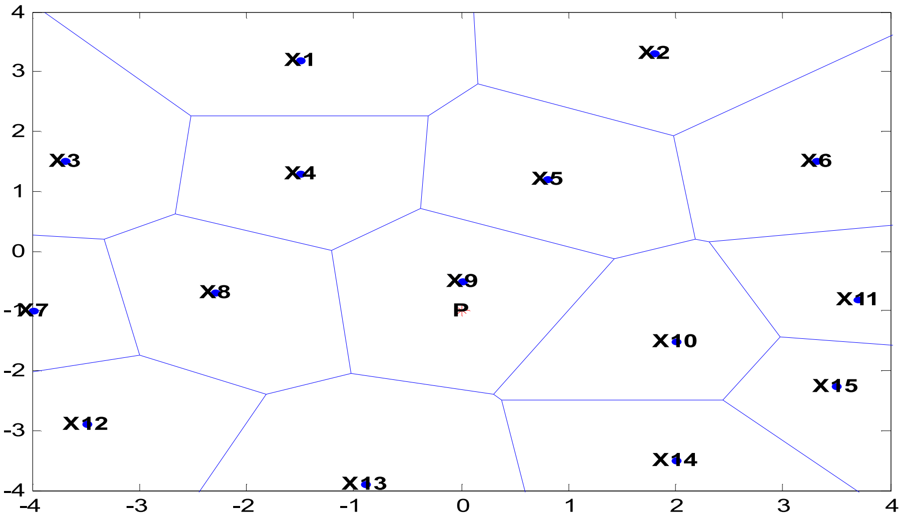

Figure [1] Voronoi Diagram

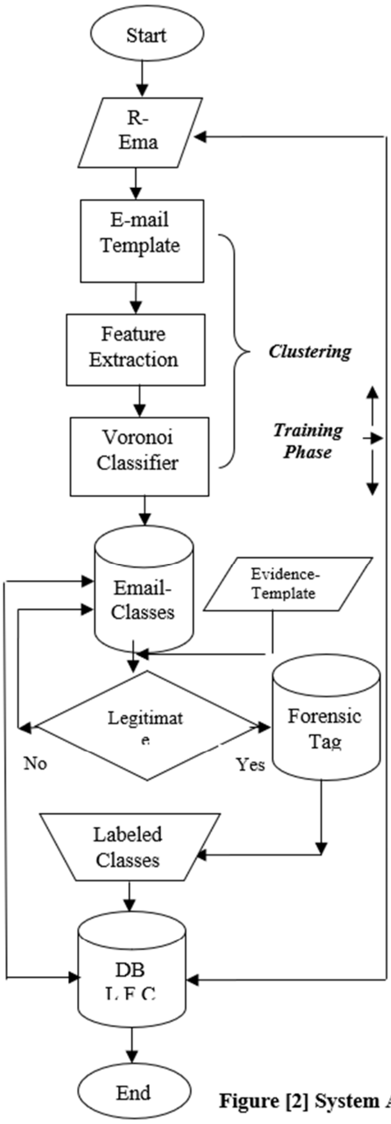

Figure [2] System Architecture

6.2. E-mail Template

Figure [2] depict proposed system block diagram, where E-mail template covers proposed E-mail features shown in table [1] which are extracted as a package data. Email candidate attributes are the first step in isolating and classifying different clusters. The data is extracted by means of the attributes mentioned in Table [1]. The e-mail ID, its source, the recipient, the subject, time of the message, the forensic features that can be indicated and the weight of these attributes are obtained in each message. The increasing the number of e-mails and the greater the size of the data will increase the accuracy of the results and this is what the conclusion in this research. What distinguishes this research is the inclusion of some processes that work on the detection of new criminal properties that can be added to the characteristics mentioned in Table 1. That the new idea and technique in this research lies in the interaction between the old and proposed new features that are discovered when training the system.

Table [1] – Template –Email

ID	Size	Origin	Destination	Subject

Time	SPF	IP	No. Forensic Tag	Mac

Phrases	Tokens	Enc.	No. of E-Mails	Weight

6.3. Features Extraction

An SPF record is a DNS record that has to be added to the DNS zone of the domain. SPF record can specify which IP addresses and/or hostnames are authorized to send email from the specific domain. [14] .

Some tags are forensic, which are suggested as a seed tags when training the system. The training system produced different additional tags with different weights. These weights are directly proportional to the strength of the indicator used in the classification, moreover during the training period; some of them may be added or deleted. Warning, kidnapping, bargaining, murder, assault, displacement, weapons, and money can all be forensic tags but with different weights. Table [2] presents the proposed initial forensic tags with its weights.

Table [2] Initial Forensic Tags

Tag	Weight	Tag	Weight	Tag	Weight
Warning	0.7	Assault	0.9	Stole	1.0
Kidnapping	1.0	disp	0.4	Attack	0.9
Bargaining	0.5	Weapon	0.8	Hit	0.5
Murder	1.0	Money	0.4	Shot	0.6

6.4. Training the System

More than twenty forensic tags discovered in the email contents during the training phase, which is processed to have suitable weights depending on the frequency of their appearance in the contents of the emails. These extracted tags are an important factor in strengthening the outer borders of the clusters. Table [3] presents some of these new tags with its weights after processing the contents of seven hundred fifty nine random messages. Weight gain is directly proportional to the frequency of tagging in email contents. The proposed research idea also relies on the extraction of influential and repetitive sections and phrases that include suggested or marked tags during the training phase. Users of e-mails repeat similar terms with similar or different tags. This repetition may be considered as further evidence that such messages have been issued by the same persons or organizations. This was adopted in this research as characteristics of classification of clusters.

Table [3] New Extracted Tags and its Weights

Tag	Weight	Tag	Weight
Spy	0.8	Terrorism	1.0
Arrested	0.7	Protect	0.6
Explosion	1.0	Explosives	1.0
Killing	1.0	Extremists	0.8
Police	0.5	Army	0.9

These tags and its morphological words were used in this phase to cover more content and to make the extraction more realistic.

```
Algorithm Extraction (EM, Phrase);
  Repeat
   I := 1;
   J := 1;
While EM[I,J] <> Φ do
 begin
EC [J] =EM [I, M]; // Read content of Email [I] as a
                          vector.
 Begin
        For k :=1 To M do
         begin
           For L :=1 to N do
             begin
               If EM [k,L] = Tag[L]
              Weight[k]:=Weight[k]+1// Increment Weight
                end;
                    end;
     While Phrase [I,J] <> Phrase[Φ,J] do
       Phrase [EM,J] := Phrase[I,J]
              end;
       Repeat
          If chkphrs [I,M] = Phrase[EM,J]
             begin
               chk[I] := chk[I]+1
               chkphrs[I,M] := chkphrs[Φ,M]
              end;
        Until chkphr[I,M]
        i := i+1 ;
        j := j+1 ;
  end;
Until EM[I,J] = Φ
```

Hhh

To determine the weight w_{kj} of term t_k in the content of email e_j represents content as a vector of weighted terms may be used. Most of the times, the standard *tfief*

To classify tags function below was used:

$$\boldsymbol{tfief(t_k,e_j) = \#(t_k,e_j).\log \frac{|\,T_r|}{\#T_r(t_k)}} \;\ldots\ldots\ldots\;(1)$$

Where $\#(t_k,e_j)$ denotes the number of times t_k occurs in e_j ,and $\#T_r(t_k)$ denotes the document frequency of term t_k, that is, the number of documents in T_r in which t_k occurs [41]. Sebastiani, Fabrizio," Machine Learning in Automated Text Categorization", Italy, Internet search, 2002. the more often a term occurs in a content, the more it is representative of its content, vice versa, the more documents a term occurs in, the less discriminating it is.

Some of ML approach relies on the availability of an initial corpus $\Omega=\{d_1,\ldots..,d_{|\Omega|}\} \in D$ of documents pre classified under $C=\{c_1,\ldots.,c|_C|\}$, that is, the values of the total function $Ø$: $D \times C \rightarrow \{T, F\}$ are known for every pair $\{d_j,c_i\} \in \Omega$ X C. A document d_j is a positive example of c_i *if* $\Phi(e_j,c_i)=T$, a negative example of c_i *if* $\Phi(e_j,c_i)=F$..

Given a corpus Ω, one may define the generality g_Ω (c_i) of a category c_i as the percentage of documents that belong to c_i :

$$\mathbf{g_\Omega(c_i)=\frac{|\{e_j \in \Omega \mid \Phi(e_j,c_i)=\mathit{T}\}|}{|\Omega|}} \;\ldots\ldots\;(2)$$

The training set generality $gTr(c_i)$ validation set generality $gva(c_i)$ and a test set generality $g_{Te}(c_i)$ of C_i may be defined in an obvious way. Therefore and for the purpose of extracting the content of emails, it is necessary first to determine the characteristics contained in that content and then to measure the weight of those characteristics in each email. The weight of those characteristics shown in Tables 2 and 3 has been invested for the purpose of classifying messages according to criminal evidence. Table 4 shows the characteristics extracted from e-mails and their weights. The drawings of the Voronoi illustrate those that have been identified by the characteristics so as to clarify the final boundary of each classification. So, to identify emails which are the target of this paper, using the Voronoi algorithm, the proposed algorithm uses the weights listed in Tables 4 and 5 to draw the boundaries of each category. Table 5 lists processed emails with minimum and maximum tags and the summation of the tags weight of each email. Table 4 and 5 presents a sample of candidate Cyber Forensic emails.

Table [4] Emails Tags and its Weights

E_m	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
E_1	0.5	1.0	0.3	0.8	0.6	0.3	0.4	0.4	0.2
E_2	0.1	0.6	0.7	0.2	0.4	0.1	0.8	0.3	0.1
E_3	0.3	0.4	0.9	0.2	0.5	0.8	1.0	0.3	0.6
E_4	0.8	0.5	0.1	0.7	0.6	0.9	0.1	0.3	0.3
E_5	0.4	0.2	0.1	0.6	0.7	0.1	0.3	0.9	0.1
E_6	0.9	0.8	0.1	0.2	0.3	0.7	0.1	0.3	0.2
E_7	1.0	0.2	0.4	0.6	0.1	0.2	0.7	0.1	0.1
E_8	0.8	0.4	0.7	0.6	0.8	0.6	0.4	0.1	0.5
E_9	0.2	0.2	0.2	0.4	0.6	0.3	0.5	0.9	1.0
E_{10}	1.0	0.2	1.0	0.3	0.3	0.7	0.6	0.2	0.6
E_{11}	0.4	0.2	0.8	0.1	0.3	0.7	0.2	0.3	0.5

Table [5] Tags weights and generality

E_m	T_{min}	T_{max}	Total	$tf_i ef(tk,e_j)$	$g_\Omega(c_i)$
E_1	0.2	1.0	4.5	7.478	0.375
E_2	0.1	0.8	4.2	7.386	0.35
E_3	0.2	0.9	5	8.537	0.416
E_4	0.1	0.9	3.9	6.659	0.325
E_5	0.1	0.9	3.4	5.805	0.283
E_6	0.1	0.9	3.6	6.147	0.3
E_7	0.1	1.0	3.4	5.650	0.283
E_8	0.1	0.8	4.9	8.617	0.408
E_9	0.2	0.9	4.3	7.342	0.358
E_{10}	0.2	1.0	4.9	8.142	0.408
E_{11}	0.1	0.8	3.8	6.683	0.316

The environment is the email space of our dataset with various properties. Using Euclidean distance formula for the optimum neighbors in x-y coordinates:

$$\boldsymbol{Dist}((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{a},\boldsymbol{b})) = \sqrt{(\boldsymbol{x}-\boldsymbol{a})^2+(\boldsymbol{y}-\boldsymbol{b})^2} \text{ …… ……(3)}$$

a and b are locations of an agent in x-y coordinates from this environment. There is a little modification between the space of the proposed research and swarm spaces in another study such that properties are standard parameters for the processed agents. So, similarities of properties are the main aim for swarm agents in the email dataset, therefore:

$$s_i^{t+1} = s_i^t + w_1 * \beta\left(\overrightarrow{m} - \overrightarrow{p_i^t}\right) + w_2 * \mu\,(\vec{n} - \overrightarrow{p_i^t}) \text{……….(4)}$$

$\overrightarrow{m}$ and $\vec{n}$ are the optimal position of the agent, while w_1 and w_2 are real weight parameters of the properties p_i^t in time *t*.

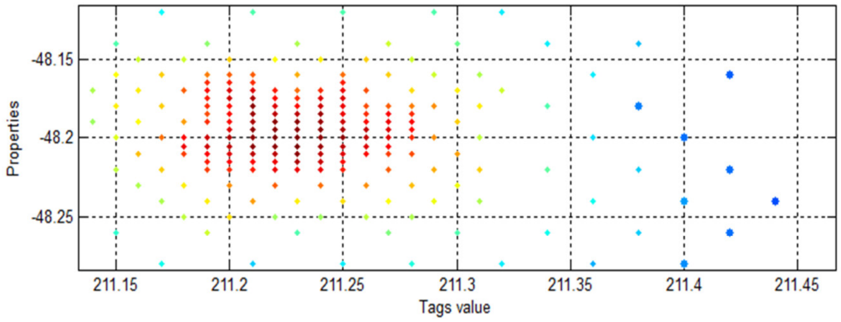

Figure [3] swarm agent optimization

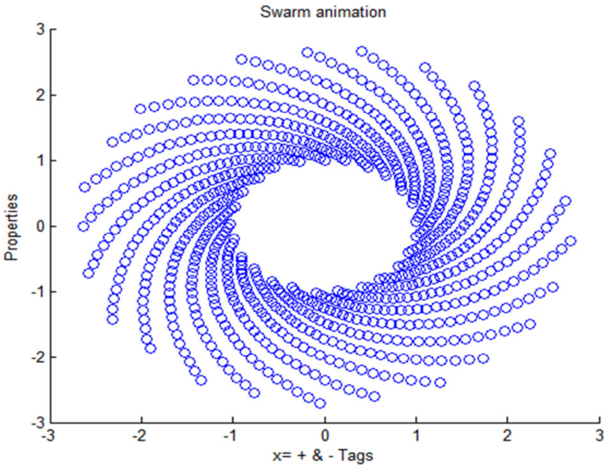

Figure [4] Swarm Animation

The movement of the swarm shown in Figure 3 depicts that all agents move around the positive tags and escape from the negative one in order to reach the general positive properties of the emails. All agents seek for emails which have forensic evidence. Note that x-coordinate represent the positive and negative tags, while y coordinate represents the general properties of the emails. So, the main goal of each individual agent is how to nearly approach the positive properties of the e-mail which classify it as a class of emails that have forensic evidence. To maximize individual agent belonging to email forensic, the similarity must be maximized:

$$\max (s_i^{t+1}) \rightarrow Ag \in \rho \text{ ....(5)}$$

where Ag: Agent and ρ arbitrary class

7. EVALUATION

One obvious method to evaluate a computational theory would be to run the program to see how well it performs. We can evaluate the results with the following metric functions:

Entropy (impurity, disorder) of a set of examples, relative to a binary classification is: -

$$Entropy\ (S) = -p_+ \log_2 (p_+) - p_- \log_2 (p_-) \text{ .... (6)}$$

Where P_+ is the proportion of positive examples in (S) and P_- is the proportion of negatives. For multiple category problems with C categories, entropy can be generalized to:

$$Entropy\ (S) = \sum_{I=1}^{C} - P_i \log_2 (P_i) \text{ .... (7)}$$

P_i is the proportion of category i examples in S.

The information gain of an attribute is the expected as reduction in entropy caused by portioning on this attribute:

$$Gain(S,A)=Entropy(S)-\sum_{v \in Values\ of\ A} \frac{|Sv|}{|S|} Entropy(S_v) \text{......(8)}$$

Where $\underline{S_v}$ is the subset of ***S*** for which attribute ***A*** has value ***v*** and the entropy of the partitioned data is calculated by weighting the entropy of each partition by its size relative to the original set.

Table [6] Entropy and Gain of the processed Emails

E_m	$g_\Omega(c_i)$	Entropy(S)	Gain(S,A)	Log (Gain)
E_1	0.375	9.333	897.209	2.952894
E_2	0.35	9.168	881.347	2.945147
E_3	0.416	11.281	1084.476	3.03522
E_4	0.325	7.829	752.625	2.876579
E_5	0.283	6.352	610.636	2.785782
E_6	0.3	7.004	673.315	2.828218
E_7	0.283	6.175	593.621	2.773509
E_8	0.408	11.431	1098.896	3.040957
E_9	0.358	9.089	873.752	2.941388
E_{10}	0.408	10.546	1013.818	3.00596
E_{11}	0.316	7.925	761.854	2.881872

Table (6) presents that E_3,E_8, and E_{10} are forensic cyber emails. These e-mails can be classified in the category of forensic messages, so that the resulting tags can be added to the previous tags and training the system on the new messages. Figure (5) depict Voronoi representation for emails data shown in table 6.

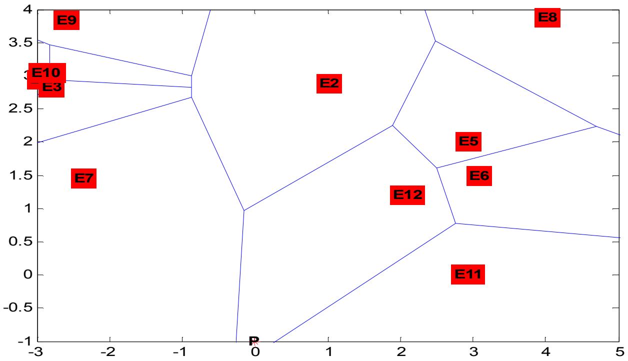

Figure [5] Voronoi graph for processed emails

While Figure (6) shows Voronoi graph for another processed emails, such that E3, E8, E9, and E10 classified in a nearest boundary and they can be in the same region of forensic emails.

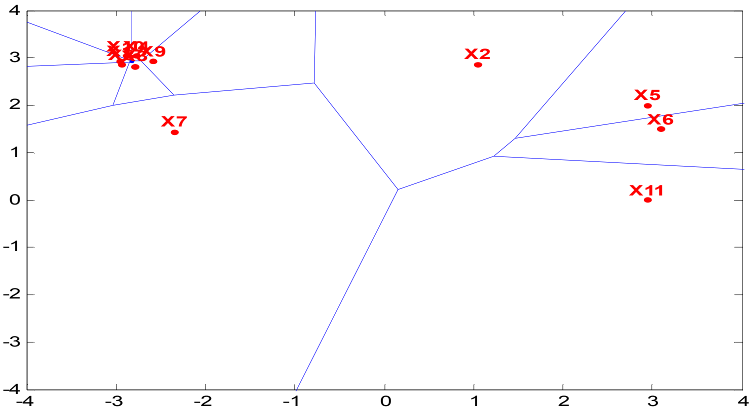

Figure [6] Voronoi graph for another processed emails

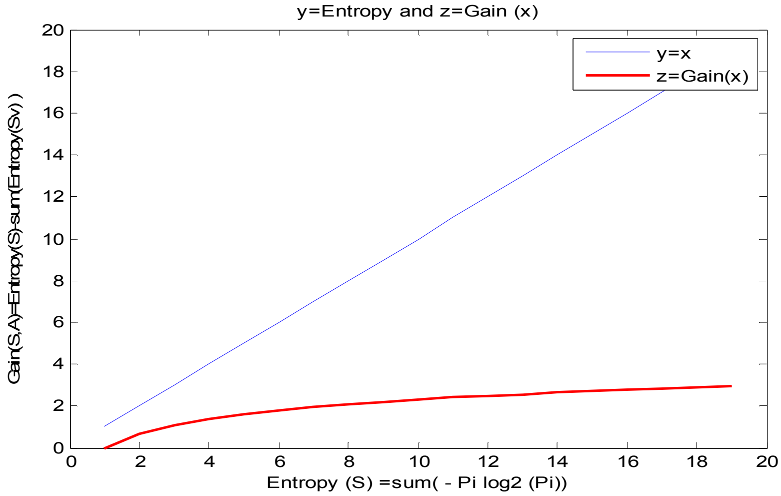

Figure [7] Entropy with Gain

Relationship between the Entropy and the Gain shown in Figure (7) depict the positive relationship between them and the accuracy of the results. The accuracy of the results increases with the increase in the size of the data.

8. Conclusions

Due to the importance of finding emails from a wide range of electronic emails that are suspected of having criminal and legal evidence of huge emails for the forensic cyber system, it has become necessary to find and motivate a technique using artificial intelligence to identify those emails. In this research, Voronoi classification was used to identify the boundary limits for each of these emails in defining space via identifying a preliminary set of attributes that could characterize this emails to which they were being referred. The proposed system was trained on a set of emails where new characteristics were added, and the results evaluated using two important evaluation criteria's: Entropy, which aims to use these characteristics for the purpose of classification, and Gain to classify a wide range of e-mails. A Limited number of emails were presented in this paper because of the inability to display all the processed emails.

ACKNOWLEDGMENT
My thanks and appreciation to all who contributed by their notes on this work. Their response and observations in Cyber Systems

References

1. Kayarkar, NIRT, and Motwani, Mining Frequent Sequences for Emails in Cyber Forensics Investigation, 2014, International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17.
2. Iqbal, Khan, Fung, and Debbabi, E-mail Authorship Verification for Forensic Investigation, 2010, ACM Digital Library.
3. Chhabra, and Bajwa, Review of E-mail System, Security Protocols and Email Forensics,2016, International Journal of Computer Science & Communication Networks,Vol5(3),201-211.
4. Khan,Nirkhi, and Dharaskar, E-mail Data Analysis for Application to Cyber Forensic Investigation using Data Mining, International Journal of Applied Information Systems (IJAIS).
5. Bemredjem, Djamel, Contributions to Cyber- Forensics: Process and Email Analysis, 2007, Thesis, Electrical and Computer Engineering, CONCORDIA University,CANADA.
6. *Laclavık,Dlugolinsky,Seleng,Kvassay,Gatial, Balogh, and Hluchy, 2011, Computing and Informatics, Vol. 30, 2011, 1001–1031, V 2011-Jan-14.*
7. Gupta,Mazumdar and Rao, 2004, Digital Forensic Analysis of E-Mails: A Trusted E-Mail Protocol, International Journal of Digital Evidence Spring 2004, Volume 2, Issue 4.
8. Hershkop ,Shlomo, Behavior-based Email Analysis with Application to Spam Detection,2006,Columbia University**.**
9. Carrizosa,Emilio, Supervised Classification and Mathematical Optimization, 2012, University of Oxford.
10. Kumar, Vipin, Data Mining and Knowledge Discovery Series, 2015, Department of Computer Science and Engineering, University of Minnesota.
11. Michie, Spiegelhalter, and Taylor, Machine Learning, Neural and Statistical Classification, 1994, Ellis Horwood Series In Artificial Intelligence.
12. Overholt, Hudas, Fiorani, Skalny, and Tucker, Dynamic Waypoint Navigation Using Voronoi Classifier Methmethods , 1994, U.S. Army Rdecom-Tardec Robotics Mobility Laboratory.
13. Aurenhammer, Klein1, Voronoi Diagrams, 1996, ACM Computing Surveys (CSUR), Volume 23 Issue 3.
14. Dmarc Analyzer, http://www.dmarcaanalyzer.com.
15. Sebastiani, Fabrizio," Machine Learning in Automated Text Categorization", 2002, Italy, Internet search.

Author-Profile

Dr. Abdulkareem Merhej Radhi is Assist. Prof. and Doctorial Philosophy in Artificial Intelligence. Supervisor of many M.Sc. students in Information Engineering Colleges rather than Science Colleges. Lecturer in Al-Nahrain University. Director of Computer Center and AvinCina for E-Learning.
Interested in Data Security, Soft Computing, Distributed Database, Engineering Analysis, Wireless Networks, Data Mining and Social Network Analysis.