doi:10.2498/cit.1000984

Suspicious E-mail Detection via Decision Tree: A Data Mining Approach

S. Appavu alias Balamurugan and Ramasamy Rajaram

Thiagarajar College of Engineering, Madurai, India

Data mining is the quest for knowledge in databases to uncover previously unimagined relationships in the data. This paper proposes to apply Decision tree in Suspected e-mail detection (e-mails about criminal activities). Deception theory suggests that deceptive writing is characterized by reduced frequency of first person pronouns and exclusive words and elevated frequency of negative emotion words and action verbs. We applied this model of deception to the set of e-mail dataset, then applied ID3 algorithm to generate the decision tree. The decision tree that is generated is used to test the e-mail as suspicious or not. In particular, we are interested in detecting fraudulent and possibly criminal activities from such data.

Keywords: data mining, deceptive theory, decision tree based classification

1. Introduction

Data mining has recently attracted considerable attention from database practitioners and researchers because of its applicability in many areas such as decision support, market strategy, financial forecasts, etc. Combining techniques from the fields like Statistics, Machine learning, Databases, etc. Data mining helps in extracting useful and invaluable information from database. Detecting unusual communication patterns in various means and channels of communications represents an important class of application directly relevant to security informatics [2].

E-mail has become one of today's standard means of communication. The large percentage of the total traffic over the internet is the e-mail. E-mail data is also growing rapidly, creating needs for automated analysis. So, to detect crime, a spectrum of techniques should be applied to discover and identify patterns and make predictions. Data mining has emerged to address problems of understanding ever-growing volumes of information for structured data, finding patterns within data that are used to develop useful knowledge.

As individuals increase their usage of electronic communication, there has been research into detecting deception in these new forms of communication. Models of deception assume that deception leaves a footprint. The work done by various researchers suggests that deceptive writing is characterized by reduced frequency of first-person pronouns and exclusive words and elevated frequency of negative emotion words and action verbs [8]. We apply this model of deception to the set of e-mail dataset and preprocess the e-mail body and to train the system we used ID3 (Iterative Dichotomiser 3) algorithm [6] to generate a decision tree that categorize the e-mail as deceptive or not.

Text classification including e-mail classification presents challenges because of large and various number of features in the data set and large number of documents. Applicability in these datasets with existing classification techniques was limited because the large number of features makes most documents undistinguishable.

In many document datasets, only a small percentage of the total features may be useful in classifying documents, and using all the features may adversely affect performance. The quality of training dataset decides the performance of both the text classification algorithms and feature selection algorithms. An ideal training document dataset for each particular category will include all the important terms and their possible distribution in the category.

To our knowledge, this is the first attempt to apply Decision tree to task of Suspicious e-mail detection (e-mails about criminal activities).

1.1. Motivation

Concern about national security has increased significantly since the terrorist attacks on 11 September 2001. The CIA, FBI and other federal agencies are actively collecting domestic and foreign intelligence to prevent future attacks. These efforts have in turn motivated us to collect the data and undertake this paper work as a challenge.

Data mining is a powerful tool that enables criminal investigators who may lack extensive training as data analyst to explore large databases quickly and efficiently. Computers can process thousands of instructions in seconds, saving precious time. In addition, installing and running software often costs less than hiring and training personnel. Also, computers are less prone to errors than human investigators. So this system helps and supports the investigators.

1.2. Organization of the Paper

The paper is organized as follows: Section 2 defines problem statement and related work in this area. Section 3 describes the proposed work and experimental results are presented in Section 4. Section 5 discusses performance measure. Finally, Section 6 concludes the paper and points out some potential future work.

2. Problem Statement and Related Work

It's hard to remember what our lives were like without e-mail. Ranking up there with the web as one of the most useful features of the Internet, billions of messages are sent each year. Though e-mail was originally developed for sending simple text messages, it has become more robust in the last few years. So, it is one possible source of data from which potential problem can be detected. Thus the problem is to find a system that identifies the deception in communication through e-mails.

One of the earlier automated deceptive detection systems, constructed from a record linkage method based on string comparators [5], was proposed by Gang Wang, Hsinchun Chen and Homa Atabakhsh. This method has a restriction that it often requires intensive computation.

Xindong Wu and Xing Xingquanzhu developed impact sensitive instance ranking method [12] to identify deception for real world data sets. This method has a restriction that the switching of attribute A_i and class C for attribute prediction AP_i, the accuracy of AP_i could be very low.

P. S. Kaila and Skillicon developed a method based on the singular value decomposition [8] to detect unusual and deceptive communication in e-mails. The problem with this approach is that it does not deal with incomplete data in an efficient and elegant way and can not incorporate new data incrementally without having to reprocess the entire matrix.

Classification is an important data mining problem. The input is a dataset of training records (also called training database), wherein each record has several attributes. Attribute with numerical domains are numerical attributes and attributes whose domains are non non-numerical are categorical attributes. There is also a distinguished attribute called the class label. This classification aims at building a concise model that can be used to predict the class label of future, unlabeled records. Many classification models including Naive Bayes, Decision tree, Support vector machine, and Neural networks have been proposed.

[13] compared a cross-experiment between 14 classification methods, including Decision tree, Naive Bayesian, Neural networks, Linear square fit, Rocchio. KNN is one of the top performers, and it performs well in scaling up to very large and noisy classification problems. [9] showed a good performance reducing the classification error by discovering temporal relations in an e-mail sequence in the form of temporal sequence patterns and embedding the discovered information into content-based learning methods. Approach to Anomalous e-mail detection is considered. [15] showed approaches to detect anomalous e-mail and involved the deployment of data mining techniques. [4] proposed a model based on the Neural network to classify personal e-mails and the use of principal component analysis as a preprocessor of NN to

reduce the data in terms of both dimensionality as well as size.

[11] and [14] developed an algorithm to reduce the feature space without sacrificing remarkable classification accuracy, but the effectiveness was based on the quality of the training dataset.

In the classification experiment for spam filtering, Decision tree showed better result than NB, NN, or SVM classifier [10].

3. Proposed Work

In this paper, we present a novel data mining based decision tree algorithm to detect e-mail concerning criminal activities. It is developed specifically for detecting deceptive communication in e-mail. The architecture of the proposed system is as follows.

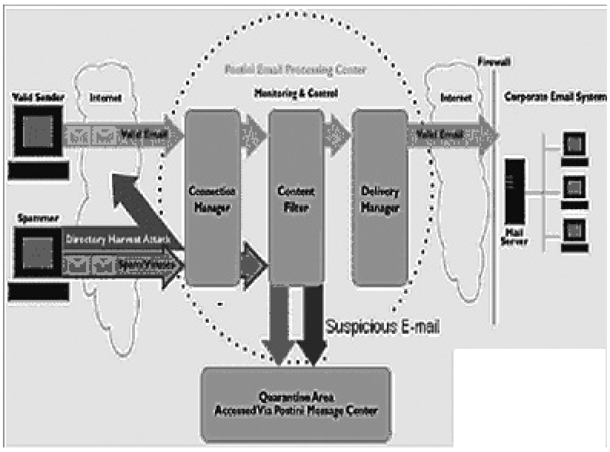


Figure 1. Proposed Suspicious e-mail detection system.

The architecture shown above is used to detect the suspicious e-mails. Connection manager is used to give the connection between the e-mail sender and the Processing center. The Content filter plays the important role i.e., it uses the preprocessing and classifying algorithm such as Decision tree, etc. to separate the suspicious e-mails. This output is delivered to the investigator with the help of Delivery manager.

The proposed method is implemented in JDK1.5 because Java is a high performance language for technical computing. In implementation, there are three parts: E-mail preprocessing, Building decision tree and Validation.

3.1. Database Used in Experiment

The Microsoft Access database is used to store the e-mail messages. The dataset contains the folder information for each of the suspicious and normal e-mail. Each message present in the folders contains the sender and receiver e-mail address, date and time, subject, body, text and some other e-mail specific technical details. We created MS Access database for the dataset to store the e-mail message, our database contains two tables. The first table contains the information of the e-mail message the sender, subject, text and other information. The second table contains the recipient's information. It contains the e-mail address of the recipient and the type (To, Cc, Bcc) in which message was sent to the recipient.

3.2. Text Classification Architecture

In Figure 2 we present a simple architecture of text classification systems. There is a pool of documents which represents the content at hand that can either be stored on disk, or could come from data streams or the web. There are standard preprocessing steps applied to this document corpus, followed by an appropriate choice of token models, representation methods, and labeling systems. Classification models are chosen to operate on train-validation-test splits, and classifiers are learned and stored.

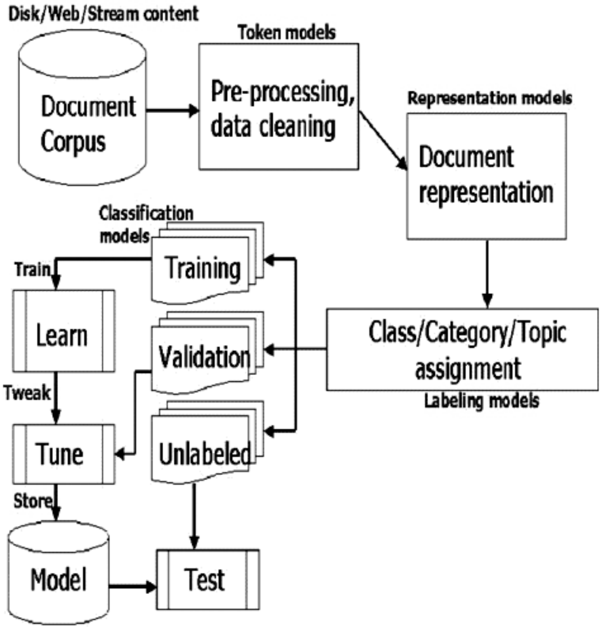


Figure 2. The standard text classification set up.

3.3. E-mail Preprocessing

E-mail preprocessing involves the process of transforming the training dataset into a representation suitable for the decision tree – ID3 (Iterative Dichotomiser 3) algorithm. This stage extracts the informational words from the data set. It consists of the following two steps:

1. Removal of non-discriminative words
2. Suffix stripping.

3.3.1. Removal of Non-discriminative Words

In e-mails, certain words are most frequent and are not discriminative of a message contents, such as prepositions, pronouns and conjunctions. Examples of such words are “we”, “that”, “and”,” this”, etc. Some widely used conversational English words, such as “I’m”, “isn’t”, “can’t”, etc. are of less importance. Elimination of these terms is performed in this step.

Based on the theory of deception, a deceptive e-mail will have highly emotional words and action verbs. So, such words are set as keywords and extracted from the input dataset and the most frequent, but less deceptive words are eliminated in this step. Examples for highly emotional words and action verbs are “lifeless”, “anger”, “kill”, “attack”, etc.

3.3.2. Suffix Stripping

Suffix stripping is a process of removing the commoner morphological and inflexional endings from words in English. Its main use is a part of a term normalization process that is usually done when setting up information retrieval systems. The Porter stemming algorithm (or ‘Porter stemmer’) is used to perform this process. Ignoring the issue of where precisely the words originate from we can say that a document is represented by a vector of words, or *terms*. Terms with a common stem will usually have similar meanings, for example:

Assassinate
Assassinated
Assassinating

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, -IONS to leave the single term assassinate. In addition, the suffix stripping process will reduce the total number of terms in the IR system, and hence reduce the size and complexity of the data in the system, which is always advantageous. Hence, those words which are extracted from the previous steps are suffix-stripped to increase their efficiency.

Unfortunately e-mails are usually very noisy and simply applying text-mining tools to them, which are usually not designed for mining from noisy data, may not bring good results.

Prior to indexing and classification, a number of preprocessing steps were performed.

1. E-mails were converted to plain-text from .mbox files.
2. Headers and HTML components were removed.
3. Body of the message was extracted.
4. The message body was tokenized into words, stop words were removed, and words were converted into lower case.

Figure 3 shows an example e-mail, which includes many typical noises (or errors) for text mining. Lines 1 and 2 are a header; lines from 4 to 8 are a signature. All of them are supposed to be irrelevant to text mining. Only line 3 is actual text content.

1. On Wednesday February 2007 13:39:42-0500, “X”
2. YYY@Domain.Com
3. Today there will be bomb blast in parliament house and the US consulates in India at 11.46 am. Stop it if you could. Cut relations with the U.S.A. Long live Osama Finladen Asadullah Alkalfi.
4.—
5. --
6. Best regards
7. X
8 --

Figure 3. Example of e-mail message.

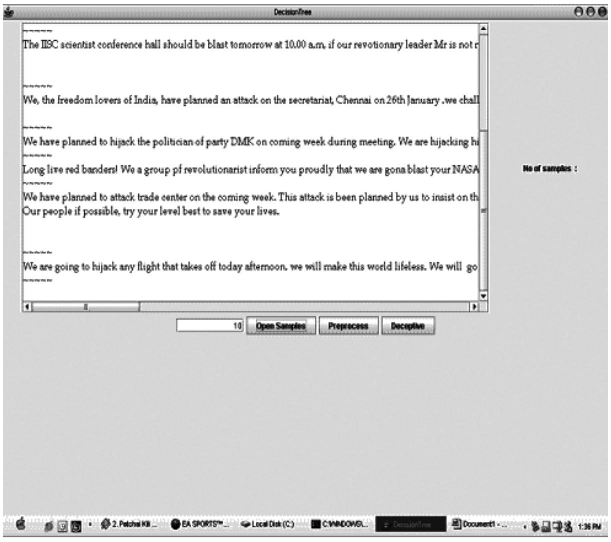
Figure 4 shows an ideal output of cleaning of the e-mail in Figure 3 within it; the non-text parts (header, signature and quotation) have been removed. The text has been normalized. Specifically, the extra line breaks have been eliminated.

1. bomb
2. blast

Figure 4. Cleaned e-mail message.

In this paper, we formalize the e-mail-cleaning problem as that of non-text data filtering and text data normalization. By 'filtering' of an e-mail we mean a process of removing the parts in the e-mail which are not needed for text mining, and by 'normalization' of an e-mail we mean a process of converting the parts necessary for text mining into texts in canonical form (like a newspaper style text).

Header, signature, quotation (in forwarded message or replied message), program code, and table are usually irrelevant for mining, and thus should be identified and removed (in a particular text mining application, however, we can retain some of them when necessary).



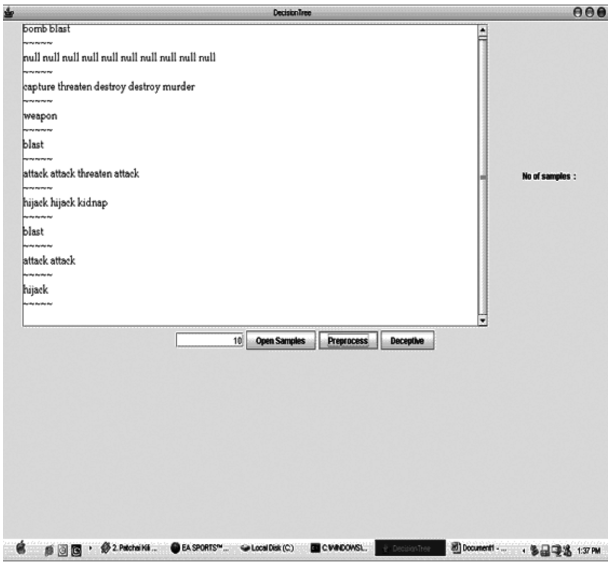
```
~~~~~
The IISC scientist conference hall should be blast tomorrow at 10.00 a.m, if our revotionary leader Mr is not r

~~~~~
We, the freedom lovers of India, have planned an attack on the secretariat, Chennai on 26th January .we chall

~~~~~
We have planned to hijack the politician of party DMK on coming week during meeting. We are hijacking hi
~~~~~
Long live red banders! We a group pf revolutionarist inform you proudly that we are gona blast your NASA
~~~~~
We have planned to attack trade center on the coming week. This attack is been planned by us to insist on th
Our people if possible, try your level best to save your lives.

~~~~~
We are going to hijack any flight that takes off today afternoon. we will make this world lifeless. We will  go
~~~~~
```

Figure 5. E-mail message before preprocessing.



```
bomb blast
~~~~~
null null null null null null null null null null
~~~~~
capture threaten destroy destroy murder
~~~~~
weapon
~~~~~
blast
~~~~~
attack attack threaten attack
~~~~~
hijack hijack kidnap
~~~~~
blast
~~~~~
attack attack
~~~~~
hijack
~~~~~
```

Figure 6. E-mail message after preprocessing.

3.4. Building the Decision Tree

We have used ID3 (Iterative Dichotomiser 3) Decision tree algorithm to classify the records. These are a few sets of e-mails used in the experiment and below is the tabulated result after preprocessing. We have shown some of the attributes in the row. The last column represents the outcome. Yes (Y) denotes the occurrence of the attribute in the e-mail and no (N) denotes the non-occurrence of the attribute in the e-mail. By using the tabulated values, the entropy and the information gain of each attribute is calculated.

E-mail	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	...	Class (D/ND)
1	1	1	0	0	0	0	0	0	0	0	0	0		D
2	0	0	1	0	0	1	0	0	0	0	1	0		D
3	1	0	1	0	0	0	0	0	0	1	0	0		D
4	1	1	1	0	0	1	0	0	0	0	0	0		D
5	1	1	0	0	0	0	0	0	0	1	0	0		D
6	0	0	0	0	0	0	0	0	0	1	1	0		D
7	1	1	0	1	0	0	0	0	0	0	0	1		D
8	0	0	0	0	0	0	0	0	0	0	0	0		ND
9	0	0	1	0	0	0	0	1	0	0	0	0		D
10	0	0	0	0	0	0	0	0	0	0	0	0		ND

A_1–BOMB, A_2–BLAST, A_3–ATTACK, A_4–THREATEN, A_5–KIDNAP, A_6–MURDER, A_7–DESTROY, A_8–WEAPONS, A_9–DANGER, $A_1$0–TERRORIST, $A_1$1–HIJACK, $A_1$2–DISASTER. 1–YES, 0–NO, D–DECEPTIVE, ND–NON-DECEPTIVE.

Table 1. Sample Training dataset.

E-mail used in the experiments: We have collected over 3000 e-mails through a brainstorming session, some of them are as follows and the first example is a real example.

Example suspicious and normal e-mail.

Suspicious e-mail	Normal e-mail
Sender: X Sub: Bomb Blast Body: Today there will be bomb blast in parliament house and the US consulates in India at 11.46 am. Stop it if you could. Cut relations with the U.S.A.long live Osama Finladen Asadullah Alkalfi.	Sender: y Sub: Hi Body: Hope ur fine! How are u & family members?

3.4.1. Algorithm for Inducing a Decision Tree from Training Samples

Input: The Training sample, samples represented by discrete value attributes; the set of candidate attributes attribute-list.

Output: A Decision tree

Method:

1) Tree is constructed in a top-down recursive divide-and-conquer manner
2) At start, all the training examples are at the root
3) Attributes are categorical (if continuous-valued, they are discredited in advance)
4) Examples are partitioned recursively based on selected attributes
5) Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

   Conditions for stopping partitioning
6) All samples for a given node belong to the same class
7) There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
8) There are no samples left

3.4.2. Proof by Induction

If S is a collection of 50 e-mails in which 38 are deceptive and 12 are non-deceptive.

$$\begin{aligned}\text{Entropy}(S) = &-(38/50)\log_2(38/50)\\ &-(12/50)\log_2(12/50)\\ &= 0.7941\end{aligned}$$

Then the information gain is calculated as follows.

There are 14 occurrences of the attribute Bomb.

$$\begin{aligned}\text{Gain}(S,\text{Bomb}) = {}& \text{Entropy}(S)\\ &-(36/50)\cdot\ \text{Entropy}(S_{\text{non-deceptive}})\\ &-(14/50)\cdot\ \text{Entropy}(S_{\text{deceptive}})\\ &= 0.7941-(36/50)\cdot(24/36,12/36)\\ &= 0.189\end{aligned}$$

Similarly, there are 10 occurrences of the attribute Attack.

$$\begin{aligned}\text{Gain}(S,\text{Attack}) = {}& \text{Entropy}(S)\\ &-(40/50)\cdot\ \text{Entropy}\ (S_{\text{non-deceptive}})\\ &-(10/50)\cdot\ \text{Entropy}\ (S_{\text{deceptive}})\\ &= 0.7941-(40/50)\cdot(28/40,12/40)\\ &= 0.133\end{aligned}$$

Similarly, there are 9 occurrences of the attribute Hijack.

$$\begin{aligned}\text{Gain}(S,\text{Hijack}) = {}& \text{Entropy}(S)\\ &-(41/50)\ \text{Entropy}\ (S_{\text{non-deceptive}})\\ &-(9/50)\cdot\ \text{Entropy}\ (S_{\text{deceptive}})\\ &= 0.7941-(41/50)\cdot(29/41,12/49)\\ &= 0.133\end{aligned}$$

$$\text{Gain}(S,\text{Blast}) = 0.0567$$

$$\text{Gain}(S,\text{Kidnap}) = 0.0499$$

Likewise, for all the attributes the information gain is calculated. The attribute which has the highest information gain becomes the root node of the tree. The attribute Bomb has got the highest information gain of about 0.189 and hence it becomes the root node. This process goes on until all the attributes are mapped in to the tree based on the sorted information gain. Following the each individual path in the tree, the rules are generated. The output of this module is Decision tree and Rules.

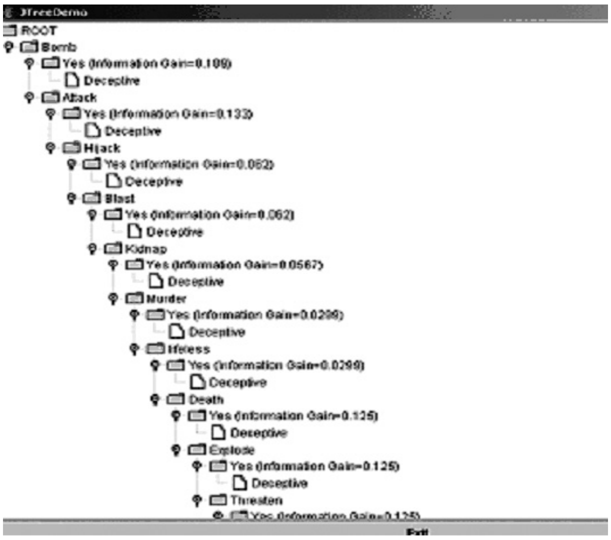
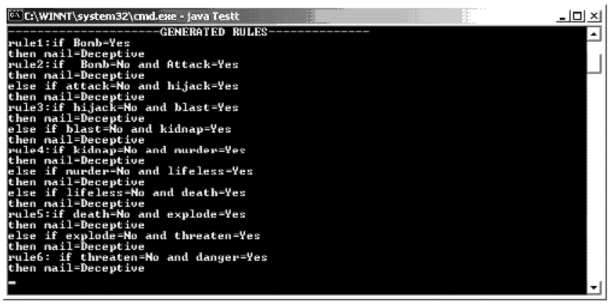

JTreeDemo
ROOT
Bomb
Yes (Information Gain=0.109)
Deceptive
Attack
Yes (Information Gain=0.133)
Deceptive
Hijack
Yes (Information Gain=0.062)
Deceptive
Blast
Yes (Information Gain=0.062)
Deceptive
Kidnap
Yes (Information Gain=0.0567)
Deceptive
Murder
Yes (Information Gain=0.0299)
Deceptive
lifeless
Yes (Information Gain=0.0299)
Deceptive
Death
Yes (Information Gain=0.125)
Deceptive
Explode
Yes (Information Gain=0.125)
Deceptive
Threaten
Yes (Information Gain=0.125)
Exit

Figure 7. Generated Decision tree.

4. Experimental Results

The application of data mining to the task of automatic suspicious e-mail classification is done; experiments were carried out on a small e-mail corpus. In order to conduct an experiment setting, different sets of 3000 e-mails are used: a mixture containing 1000 suspicious e-mails and 2000 normal e-mails. The system was trained with the training dataset and the information gain and the entropy were calculated. When the training process was finished, the best quality rules were taken as the final classification rules. Some of the rules generated by the Decision tree based classifications are:



```
--------------------GENERATED RULES--------------
rule1:if Bomb=Yes
then mail=Deceptive
rule2:if  Bomb=No and Attack=Yes
then mail=Deceptive
else if attack=No and hijack=Yes
then mail=Deceptive
rule3:if hijack=No and blast=Yes
then mail=Deceptive
else if blast=No and kidnap=Yes
then mail=Deceptive
rule4:if kidnap=No and murder=Yes
then mail=Deceptive
else if murder=No and lifeless=Yes
then mail=Deceptive
else if lifeless=No and death=Yes
then mail=Deceptive
rule5:if death=No and explode=Yes
then mail=Deceptive
else if explode=No and threaten=Yes
then mail=Deceptive
rule6: if threaten=No and danger=Yes
then mail=Deceptive
```

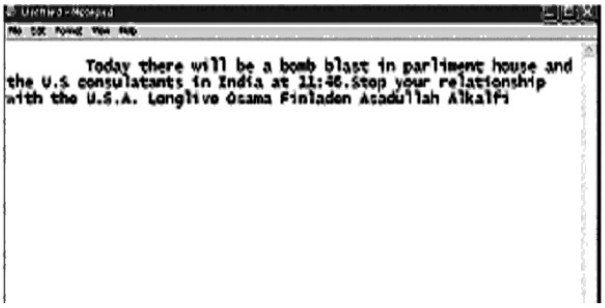
Rule 1: If Bomb=Yes then mail=Deceptive

Rule 2: If Bomb=No and Attack=Yes then mail=Deceptive else if Attack=No and Blast=yes the mail=Deceptive

Rule 3: If Blast=No and Hijack=Yes then mail=deceptive else if Hijack=No and Murder=Yes then mail=Deceptive

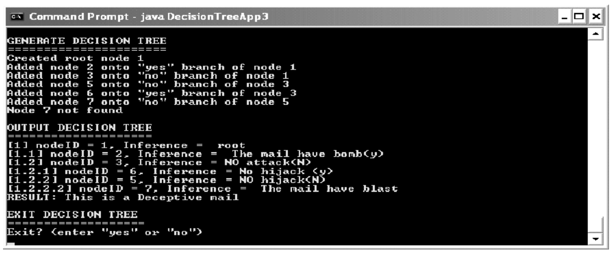
Rule 4: If Murder=No and Death=Yes then mail=Deceptive else if Death=No and Terrorist=Yes then mail=Deceptive else if Terrorist=No and Destroy=yes then mail=Deceptive

This will be the input to the testing stage:



```
            Today there will be a bomb blast in parliment house and
the U.S consulatants in India at 11:40.Stop your relationship
with the U.S.A. Longlive Osama Finladen Asadullah Alkalfi
```

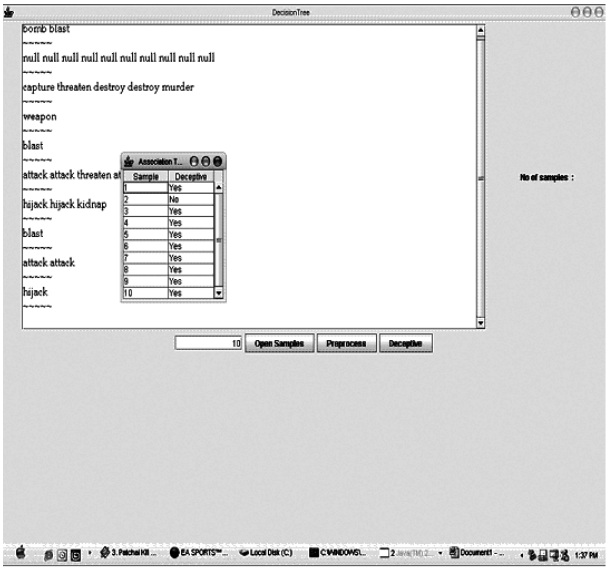
This is the output that is obtained during the execution:



```
GENERATE DECISION TREE
======================
Created root node 1
Added node 2 onto "yes" branch of node 1
Added node 3 onto "no" branch of node 1
Added node 5 onto "no" branch of node 3
Added node 6 onto "yes" branch of node 3
Added node 7 onto "no" branch of node 5
Node 7 not found

OUTPUT DECISION TREE
====================
[1] nodeID = 1, Inference =  root
[1.1] nodeID = 2, Inference =  The mail have bomb(y)
[1.2] nodeID = 3, Inference = NO attack(N)
[1.2.1] nodeID = 6, Inference = No hijack (y)
[1.2.2] nodeID = 5, Inference = NO hijack(N)
[1.2.2.2] nodeID = 7, Inference =  The mail have blast
RESULT: This is a Deceptive mail

EXIT DECISION TREE
==================
Exit? (enter "yes" or "no")
```



```
bomb blast
~~~~~
null null null null null null null null null null
~~~~~
capture threaten destroy destroy murder
~~~~~
weapon
~~~~~
blast
~~~~~
attack attack threaten at
~~~~~
hijack hijack kidnap
~~~~~
blast
~~~~~
attack attack
~~~~~
hijack
~~~~~
```

Sample	Deceptive
1	Yes
2	No
3	Yes
4	Yes
5	Yes
6	Yes
7	Yes
8	Yes
9	Yes
10	Yes

Figure 8. Proposed framework for suspicious e-mail detection.

5. Performance Measure

To evaluate the classifier on testing data, we defined an accuracy measure as follows.

Accuracy(%)=correctly_classified_emails
/Total_emails∗100

Also, precision and recall were used as the metrics for evaluating the performance of each e-mail classification approach.

A. Effect of dataset on performance

An experiment measuring the performance against the size of dataset was conducted using dataset of different sizes. For example, in case of 1000 dataset, accuracy was 96.25% using DT classifier. Decision tree classifier supports over 95% of classification accuracy for more than 1000 dataset.

Data size	DT
500	93.40
1000	96.25
1500	95.17
2000	95.38

Table 2. Classification result based on Data size.

B. Effect of feature size on performance

The other experiment measuring the performance against the feature size was conducted using different features. For example, in case of 20 features, accuracy was 93.91% using DT classifier. The most frequent words in suspicious e-mail were selected as features. Generally, the result of classification was increased for all classification methods according the feature size increased. Decision tree classifier supported over 94% of classification accuracy for more than 30 feature size.

Feature size	DT
10	91.64
20	93.91
30	94.46
40	94.16
50	94.64

Table 3. Classification result based on Feature size.

6. Conclusions and Further Work

E-mail is an important means for communication. It is a possible source of data from which potential problem can be detected. In this paper, we have employed decision tree-based classification approach to detect e-mails in relation to criminal activities. All the e-mails were classified as suspicious (1) or not (0). From this experiment, we can find that a simple decision tree classifier can provide better classification result for suspicious e-mail detection. In the near future, we plan to incorporate other techniques such as different ways of feature selection, and classification using other method. One major advantage of the decision tree-based classifier is that it doesn't assume that terms are independent and its training is relatively fast. Furthermore, the rules are human understandable and easy to be maintained. The proposed work will be helpful for identifying the deceptive e-mail and will also assist the investigators to get the information in time to take effective actions to reduce criminal activities. In the future, we would add following features to the paper: automatic reply to the incoming e-mails that are found deceptive and enabling our application to work in mobile environment.

A problem we faced when trying to test out new ideas dealing with e-mail systems was an inherent limitation of the available data. Because we only have access to our own data, our results and experiments no doubt reflects some bias. Much of the work published in the e-mail classification domain also suffers from the fact that it tries to reach general conclusion using very small data sets collected on a local scale.

7. Appendix

We have collected over 3000 e-mails through a brainstorming session, some of them are as follows and the first example is a real example.

Live example:

Today there will be bomb blast in parliament house and the US consulates in India at 11.46 am. Stop it if you could. Cut relations with the U.S.A. Long live Osama Finladen Asadullah Alkalfi.

Sample training dataset

We are a band of patriot who has currently captured a Brahmas and a nuclear warhead. We threaten to destroy the parliament building in session if any is held. With captures technology and our expertise. We have started mass production of this ICMS.we can destroy any place, any time since we have many strategically placed base stations, throughout the planet.

We, the freedom lovers of India, have planned an attack on the secretariat, Chennai on 26th January. We challenge you for the attack and stop it if you can. This is not just a threaten mail. So get ready for the attack.

We have planned to attack trade center on the coming week. This attack is been planned by us to insist on the freedom of our people. We are ready to loss our lives for the sake of Our people if possible, try your level best to save your lives.

The lists of supporters of RAM temple construction have been found region wise. These coverts are going to face drastic consequences which no one had ever seen in their life. It may include your name. So roll back your efforts otherwise?????

One of the key defense research labs of the national security /national defense agent will be blown within the next 120 hours.

We are going to attack the US embassy. There will be a terrorist attack or a bomb scare or anthrax spread. Beware of it. We are going to demolish the US.consultate Building beware .if possible try to stop it.

We, the terrorist society have planned to do a bomb blast in the Indian airline plane today. We don't like this government. We do this act as criticism to this government.

We are going to attack the parliament on 28-1-2006. The bomb will blast at any time .we plan to kill the leaders. The attack may be on 26-1-06 to demolish. The republic day parades. It is not possible to prevent the disaster if possible save them.

A bomb has been placed in Tajmahal. It may explode on today .save your people if you wish to.

References

[1] S. Appavu alias Balamurugan, R. Rajaram, S. Senthamarai Kannan, A Novel Data mining approach to Detect Deceptive Communication in Email Text. *Proceedings of the National Conference on Advanced Computing*, MIT, Chennai, India, (2007), pp. 179–188.

[2] A. Badia, M. M. Kantardzic, Link Analysis Tools for Intelligence and Counterterrorism. *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, Atlanta, GA, (2005), pp. 49–59.

[3] W. Cohen, Learning rules that Classify Email. *In proc.of the AAAI Spring Symposium on Machine Learning in Information Access*, (1996).

[4] B. Cui, A. Mondal, J. Shen, G. Cong, K. Tan, On Effective Email Classification via Neural Networks. *In Proc. of DEXA*, (2005), pp. 85–94.

[5] G. Wang, H. Chen, H. Atabakhsh, Automatically Detecting Deceptive Criminal Identity. *Comm. ACM*, (2004), pp. 70–76.

[6] J. Han, M. Kamber, Data Mining Concepts and Techniques. *Morgan Kaufmann Publishers*, 2004.

[7] J. Tang, H. Li, Y. Cao, Z. Tang, Email Data Cleaning. *Proceedings of KDD*, Chicago, USA, (2005).

[8] P. S. Keila, D. B. Skillicorn, Detecting Unusual and Deceptive Communication in Email. *Technical Reports*, (June 2005).

[9] S. Kiritchenko, S. Matwin, S. Abu-Hakima, Email Classification with Temporal Features. *Intelligent Information Systems*, (2004), pp. 523–533.

[10] S. Youn, D. McLeod, A Comparative Study for Email Classification. *Proceedings of International Joint Conferences on Computer, Information, System Sciences and Engineering*, Bridgeport, CT, (2006).

[11] S. Shankar, G. karypis, Weight Adjustment Schemes for a Centroid based Classifier. *Computer Science Technical Report TR00-35*, (2000).

[12] X. Wu, X. Xingquanzhu, Data Acquistation with Active and Impact Sensitive Instance Selection. *16th IEEE interactive conference*, (2004).

[13] Y. Yang, An Evaluation of Statistical Approaches to Text Categorization. *Journal of Information Retrieval*, Vol. 1, No. 1/2, (1999), pp. 67–88.

[14] Y. Yang, J. Pedersen, A Comparative study on Feature selection in Text Categorization. *In ICML*, (1997), pp. 412–420.

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

Received: November, 2006
Accepted: February, 2006

Contact addresses:

S. Appavu alias Balamurugan
Faculty, Dept. of Information Technology
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
e-mail: sbit@tce.edu

Ramasamy Rajaram
Prof. & Head, Dept. of Computer Science
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
e-mail: hodcse@tce.edu

S. Appavu alias Balamurugan received the B.E. degree in Electronics and Communication Engineering from Mohamed Sathak Engineering College, Kilakarai, in 2001 and the Master's in Computer Science and Engineering from the University of Madras, Chennai. He is currently pursuing Ph.D. in Information and Communication Engineering from Anna University, Chennai. He has been working in the area of databases focusing on data mining, data warehousing and cyber security.

Dr. Ramasamy Rajaram, Head of the Department of Computer Science, Thiagarajar College of Engineering, Madurai, secured his B.E. E.E.E. from Madras University in 1966, M Tech in E.E.E. from IIT Kharagpur in 1971 and Ph.D. from Madurai Kamaraj University in 1979. He currently teaches and guides research in data mining, genetic algorithms and cyber security.