# Association Rule Mining for Suspicious Email Detection: A Data Mining Approach

S.Appavu alias Balamurugan, Aravind, Athiappan, Bharathiraja, Muthu Pandian and Dr.R.Rajaram

**Abstract**-Email has been an efficient and popular communication mechanism as the number of internet user's increase. In many security informatics applications it is important to detect deceptive communication in email. This paper proposes to apply Association Rule Mining for Suspected Email Detection.(Emails 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 apply this model of deception to the set of Email dataset, then applied Apriori algorithm to generate the rules .The rules generated are used to test the email as deceptive or not. In particular we are interested in detecting emails about criminal activities. After classification we must be able to differentiate the emails giving information about past criminal activities(Informative email) and those acting as alerts(warnings) for the future criminal activities. This differentiation is done using the features considering the tense used in the emails. Experimental results show that simple Associative classifier provides promising detection rates.

**Index Terms**- Data Mining, Deceptive Theory, Association Rule Mining, Apriori algorithm, Tense.

## 1. INTRODUCTION

E-mail has become one of today's standard means of communication. The large percentage of the total traffic over the internet is the email. Email 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. Work done by various researches 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 [KS05]. We apply this model of deception to the set of E-mail dataset and preprocess the email body and to train the system we used Apriori algorithm to generate a classifier that categorize the email as deceptive or not.

### 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 data's 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 personality. Computers are also less prone to errors than human investigators. So this system helps and supports the investigators.

To our knowledge, this is the first attempt to apply Association rule mining to task of suspicious Email Detection (Emails about criminal activities). The reason is that we have included the concept of extracting the informative emails using the tense (Past tense) of the verbs used in the emails. Apart from the informative emails, other emails are considered as the alerting emails for the future occurrences of hazard activities.

The remainder of this paper is organized as follows: Section 2 gives an overview of Problem Statement & related work in Email classification. In section 3 we introduce our new Suspicious Email detection approach. Experimental results are described in section 4 .We summarize our research and discuss some future work direction in section 5.

## 2. PROBLEM STATEMENTS AND RELATED WORK

S.Appavu alias Balamurugan is with the Dept of Information Technology, Thiagarajar College of Engineering, Madurai-15, Tamilnadu, India.E-mail: app_s@yahoo.com

Dr.R.Rajaram is with the Dept of Computer Science, Thiagarajar College of Engineering, Madurai-15, Tamilnadu, India.

1-4244-1330-3/07/$25.00 ©2007 IEEE.

It's hard to remember what our lives were like without email. Ranking up there with the web as one of the most useful features of the Internet, billions of messages are sent each year. Though email 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 emails. Even after classification of deceptive emails we must be able to differentiate the informative emails from the alerting emails. We refer to informative emails as those giving details about the already happened hazardous events and the alert emails are those which remain us to prevent those hazard events to occur in the fore coming days.

**Example of suspicious and normal email.**

| Suspicious Email | Normal Email |
|---|---|
| Sender: X<br>Sub: Bomb Blast<br>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<br>Sub: Hi<br>Body: Hope ur fine! How are u & family members? |

**Example of classifying Suspicious into Alert and informative email:**

| Alert Email | Informative Email |
|---|---|
| Sender: X<br>Sub: Bomb Blast<br>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<br>Sub: WTC Attacked<br>Body: The World Trade Center was attacked on 9/11/01 by Osama Bin Laden and his followers. |

The informative emails provides us with the data about the past historical criminal activities by enhancing some common sense to us such as in the example shown above we came to know that these types of email will never have any consequences in future.

The alert emails were identified using the deceptive theory and the future tense verbs used in the emails. By which the security enforcing methods can be strengthened. Also we can prevent the occurrences of future attacks

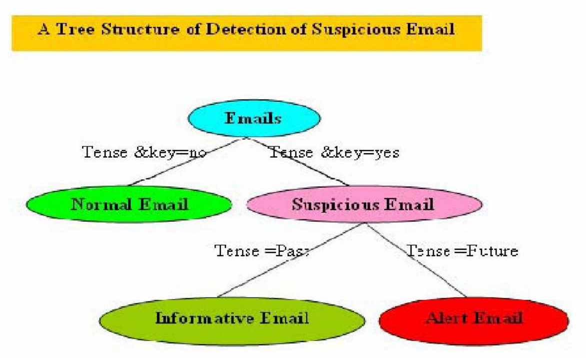


**Fig. 1. A Tree Structure of Detection of Suspicious Email**

Many techniques such as Naïve bayes [LEW98,CDAR97,ABSS00], Nearest Neighbor [GL97],Support Vector Machines [JOA 98], Regression [YC94],Decision Trees[ADW98],TF-IDF Style Classifiers [SM83,BS95,ROC71] and Association classifiers [LHM98,WZL99] have been developed for text classification.

[COH 96] compares results for email classification of a new rule induction method and adaptation of Rocchio's relevance feed back algorithm [ROC71] in [ILA95]. [SDHH98] employs Naïve Bayes Classifier to filter junk email.[BOO98] uses a combination of nearest neighbor and TF-IDF approaches .Naive Bayes classifier is used for classifying email in to multiple categories in [REN00].Support Vector machines approach is implemented for email authorship classification in [VE00]. A comparison of binary classification using Naïve bayes and decision trees [QUI93] approaches is performed in [DLW00].TF-IDF style classifier defined in [BS95] is implemented in [SK006] and is extended for incremental case in [SK00A].Approach to Anomalous email detection is considered [ZD] showed approaches to detect Anomalous email involves the deployment of data mining techniques. [CMSCT] Proposed a model based on the Neural Network to classify personal emails 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.

Using association rules for classification was first introduced in [LHM98] and further developed in [WZL99,MW99,WZH00,LI01].Classification based on Association rule (CBA) was introduced in [LHM98] and Multiple Association rule (CMAR) introduced in [LHM98,LI01].[KS05] proposed a method based on the singular value Decomposition to detect unusual and Deceptive communication in emails. The problem with this approach is that not deals with incomplete data in an efficient and elegant way and can not able to incorporate new data incrementally without having to reprocess the entire matrix.

No work is known to exist that would test Associative classification specifically to detect email concerning criminal activities.

## 3. THE PROPOSED WORK

In this paper, we present an association rule mining algorithm (Apriori algorithm) to detect suspicious email and the further classification into the alert and informative emails. It is developed specifically for detect unusual and deceptive communication in email. The proposed method is implemented using the java language. In implementation, there are three parts: Email Preprocessing, Building the associative classifier and validation.Fig.2. Shows a general framework of the Associative classifier construction.

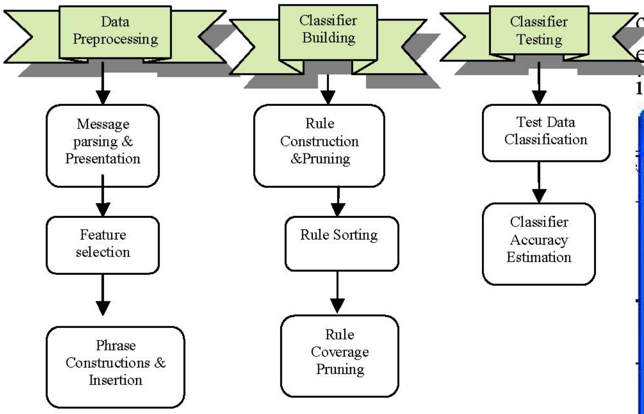


**Fig.2.Classifier Construction framework**

### 3.1. E-Mail Preprocessing

Email preprocessing involves the process of transforming email messages into a representation suitable for the Apriori algorithm.

### 3.1.1 Email Data Extraction and presentation

This stage extracts highly emotional and actioned words from a message body and subject fields these individual words are helpful in differentiating the suspicious emails from the normal emails. Then the tense of the separated suspicious emails were examined to differentiate the alert and informative emails. It consists of the following four steps: Text term extraction, lexical analysis, stop word removal and stemming.

**Text term Extraction**

Extract valid text terms from a document, performing stop list word removal and stemming.

**Lexical Analysis**

Lexical analysis is the process of converting an input steam of characters in to a stream of words or tokens. The lexical analysis phase produces candidate terms that are further checked and retained if they are not in a stop list.

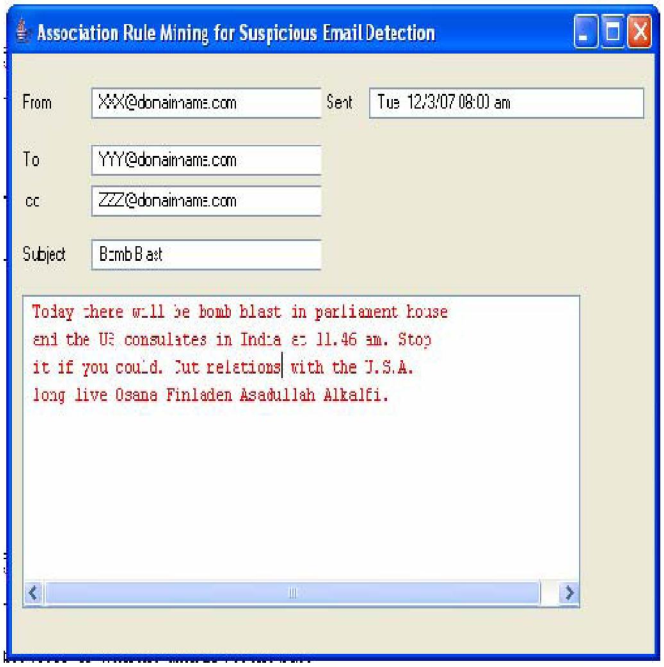
**Stop Word Removal**

Stop list is a list of words that are most frequent in a text corpus and are not discriminative of a message contents, such as prepositions, pronouns and conjunction. Examples of stop words are “the”, “and”, “about”, etc.

**Stemming**

Stemming is the process of suffix removal to generate word stems. Although not always absolutely true, terms like “bomb”, and “bombing” do not make big difference for the purpose of distinguishing messages containing trip bombing, for example, and can all be replaced by their stem “bomb”.

Consider the email message on Fig 3.1. and Fig 3.2.The body of the first email content is given in red color to denote the reader that it is a suspicious email. The body of the second email content shows it is not a suspicious.



Association Rule Mining for Suspicious Email Detection

From XXX@domainname.com Sent Tue 12/3/07 08:00 am

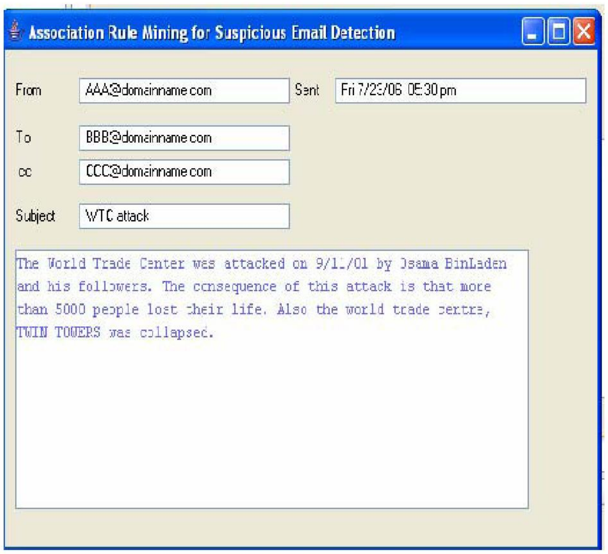
To YYY@domainname.com

cc ZZZ@domainname.com

Subject Bomb Blast

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.

**Fig.3.1. Semi Structured data (alert email)**



From: AAA@domainname.com  Sent: Fri 7/23/06 05:30 pm

To: BBB@domainname.com

cc: CCC@domainname.com

Subject: WTC attack

```
The World Trade Center was attacked on 9/11/01 by Osama BinLaden
and his followers. The consequence of this attack is that more
than 5000 people lost their life. Also the world trade centre,
TWIN TOWERS was collapsed.
```

**Fig.3.2. Semi structured data (informative email)**

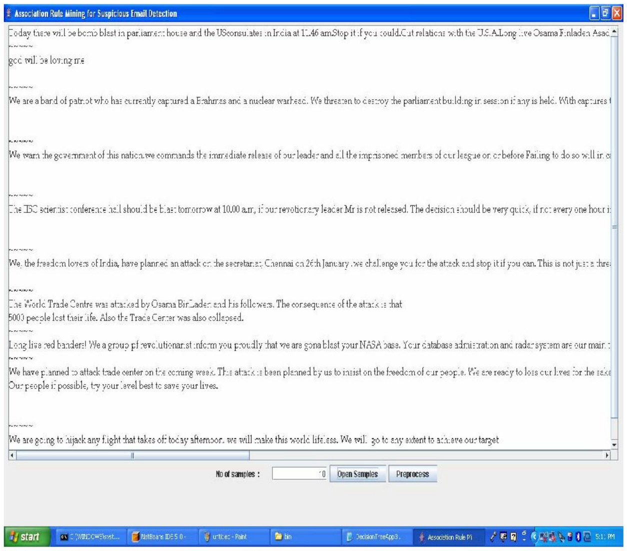
### 3.1.2. Feature Selection

Based on the theory of deception a deceptive email will have highly emotional words and action verbs. So, such words are set as keywords and extracted from the input dataset. Example for highly emotional words and action verbs are “lifeless”, “anger”, “kill”, “attack”, etc.The future tense denoting keywords such as will, shall, may, might, should, can, could, would are used to indicate that the suspicious email is of the type alert. The past tense denoting keywords such as was, were, etc are used to indicate that the suspicious email is of the informative type.

Prior to classification, a number of preprocessing steps were performed

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

After the above mentioned steps the email is given to the preprocessing program and the Fig.4. gives the view of the email before preprocessing.

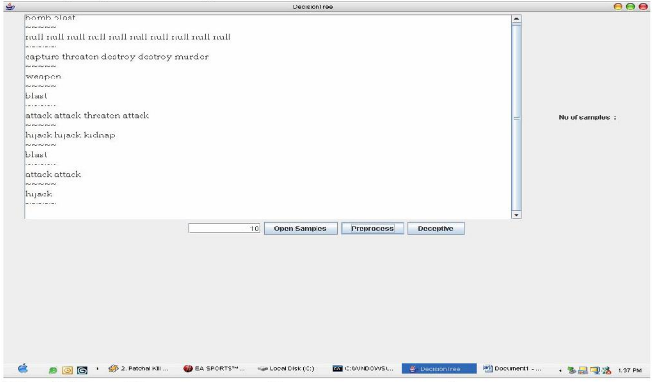


```
Today there will be bomb blast in parliament house and the USconsulates in India at 11.46 am.Stop it if you could.Cut relations with the U.S.A.Long live Osama Binladen Asad
~~~~~
god will be loving me

~~~~~
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 t


~~~~~
We warn the government of this nation.we commands the immediate release of our leader and all the imprisoned members of our league on or before Failing to do so will in ca


~~~~~
The ISC scientist conference hall should be blast tomorrow at 10.00 a.m, if our revotionary leader Mr is not released. The decision should be very quick, if not every one hour i


~~~~~
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 thre

~~~~~
The World Trade Centre was attacked by Osama BinLaden and his followers. The consequence of the attack is that
5000 people lost their life. Also the Trade Center was also collapsed.
~~~~~
Long live red banders! We a group pf revolutionarist inform you proudly that we are gona blast your NASA base. Your database administration and radar system are our main t
~~~~~
We have planned to attack trade center on the coming week. The attack is been planned by us to insist on the freedom of our people. We are ready to loss our lives for the sake
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 to any extent to achieve our target
```

No of samples : 10

**Fig.4.Email message before preprocessing**

The email after preprocessing is of the form that removes space and extra characters and displays only the keywords. The Fig.5. gives the view of the email after 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
~~~~~~
```

No of samples : 10

**Fig.5.Email message after preprocessing**

The output after the preprocessing is in the table format in which the attributes are given as the table headers and the records are given in column. The class attribute is to detect either informative, alert or normal email.

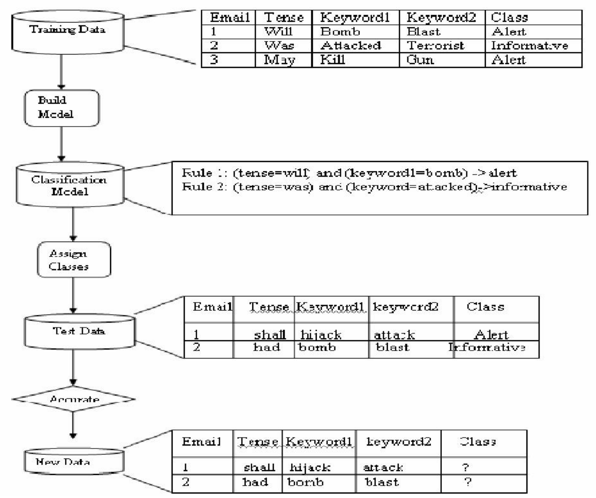
| Email | Tense | Bomb | Blast | Terrorist | Attack | Threaten | Class |
|---|---|---|---|---|---|---|---|
| 1 | past | y | y | y | y | n | informative |
| 2 | past | n | n | y | y | y | informative |
| 3 | present | y | y | y | y | n | alert |
| 4 | future | n | y | n | y | y | alert |
| 5 | past | n | n | n | n | n | normal |
| 6 | present | y | y | y | n | n | alert |
| 7 | past | n | n | n | n | y | informative |
| 8 | past | y | y | y | y | y | informative |
| 9 | future | n | y | n | y | y | alert |
| 10 | future | y | n | y | n | y | alert |

**Table 1: Final output of preprocessing**

### 3.2. Building the Associative Classifier

### 3.2.1. Email Classification Process:

Email Classification is the process of finding a set of models (or functions) that describes and distinguish data classes and concepts, for the purpose of being able to use the model to predict class of objects whose label is unknown.



| Email | Tense | Keyword1 | Keyword2 | Class |
|---|---|---|---|---|
| 1 | Will | Bomb | Blast | Alert |
| 2 | Was | Attacked | Terrorist | Informative |
| 3 | May | Kill | Gun | Alert |

Rule 1: (tense=will) and (keyword1=bomb) -> alert
Rule 2: (tense=was) and (keyword=attacked)->informative

| Email | Tense | Keyword1 | keyword2 | Class |
|---|---|---|---|---|
| 1 | shall | hijack | attack | Alert |
| 2 | had | bomb | blast | Informative |

| Email | Tense | Keyword1 | keyword2 | Class |
|---|---|---|---|---|
| 1 | shall | hijack | attack | ? |
| 2 | had | bomb | blast | ? |

**Fig.6. Classification Process**

Fig 6 shows a general framework of the classification process. In the example, objects correspond to email messages and object class labels correspond to message type. Every email message contains two terms and a Tense, that are used to predict a email is suspicious or not. The training set contains three email messages. For each message we recorded two terms or more (it can be a single word or a Tense), Term1, Term2 and a Tense. And class label was pre assigned to each message manually.

Let the following be a simplified informal algorithm, for classification model construction. If one or several attributes $a_i$ occur together in more than one transaction assigned the same topic T, then output a rule $a_i$->T.In the first step we built a classification model. The training data contains two transactions of class Alert Email that have keyword Bomb/ Blast/kill and a tense "will/may" in them, one transaction of class informative email that have keyword Attacked/Terrorist and a tense "was". Applying the Apriori algorithm we obtain a model containing two rules shown as the classification model. In the second step the model just built is tested using test data containing two transactions. If accuracy is measured as a percentage of messages correctly classified, If accuracy is not satisfactory then one or several steps of the classifier need to be modified.

We have used Apriori algorithm to classify the emails. These are the few set of emails used in the experiment and below is the tabulated result after preprocessing.

**Email 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,

**An 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.

| Email ID | Items or keywords |
|---|---|
| 1 | Will,Bomb,Blast,terrorist,attack |
| 2 | May, Terrorist, attack,threaten |
| 3 | Was, Blast, attack, threaten, |
| 4 | -------- |
| 5 | Bomb, blast, terrorist,attack,threaten |
| 6 | Can, Bomb, terrorist, threaten |
| 7 | Was, Hijack, murder |
| 8 | Could, Attack, Disaster |
| 9 | Was, Terrorist, Bomb, blast |
| 10 | Will, Attack, hijack, murder |
| 11 | Might, Attack, bomb, blast, kill, demolish, disaster |
| 12 | Will, Attack, hijack, murder |
| 13 | Was, Terrorist, Bomb, blast |
| 14 | Shall, Attack, bomb, blast, kill, demolish, disaster |
| 15 | Will, Hijack, murder |

**Table. 2. Sample Feature Selection from email**

### 3.2.2. Apriori Algorithm for Suspicious Email Detection

Association Rule mining searches for interesting association or correlation relationships among items in a given large data set. We model email messages as transaction where items are words or phrases from the email. After preprocessing a email message, by eliminating stop words and stemming, emails are represented by sets of cleaned words di= $\{t_1 \ldots t_n\}$ as well as category to which they belong $C_j$. The Apriori algorithm is used for mining frequent item-sets in transactional databases to find frequent sets of words in the emails of the training set. Given the frequent sets of words and topical category assigned to the transaction from which they were extracted association rules are deduced with constraints on the antecedent and consequent of the rules such that the antecedent always contains words while the consequent is exclusively a topical category.

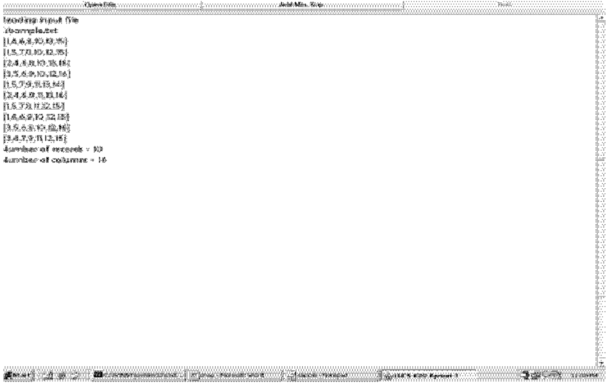
The input to the association rule generating program gets the values in numerical order hence we are assigning the values to the attribute as given in the table.

**Table 3: Assigned value for each attribute**

Suppose and confidence is the two measures that are used in association rule mining. Support can be defined as fraction of transaction that contains both X & Y.Confidence measure how often items in Y appear in transaction that contain X.

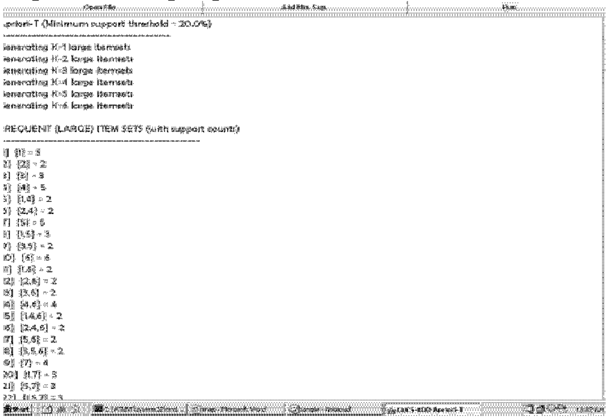
| Attribute | Value 1 | Value 2 | Value 3 |
|---|---|---|---|
| Tense | Past =1 | Present =2 | Future =3 |
| Bomb | Yes =4 | No =5 | ------------ |
| Blast | Yes =6 | No =7 | ------------ |
| Terrorist | Yes =8 | No =9 | ------------ |
| Attack | Yes =10 | No =11 | ------------ |
| Threaten | Yes =12 | No =13 | ------------ |
| Class | Normal =14 | Informative =15 | Alert =16 |

Market Basket analysis is possibly the largest application for algorithms that discover Association Rules. The application of Association Rules is not restricted to market basket data. This paper attempt to apply the algorithm for suspicious email detection as informative or alert emails. The algorithm produces a list of Itemsets. Each itemset is a combination of attribute values. For example, Itemsets for the suspected informative email could include {Tense=past, Attack= Yes, Bomb = Yes} and Item sets for the suspected alert email could include {Tense=future, Attack= Yes, Bomb = Yes}. The support shows how many cases have the Itemset values {Tense =past, Attack = Yes, Bomb = Y, Email =Suspicious informative email} and the confidence shows the likelihood of Email = Suspicious or Deceptive for a case having Attack = Y and Bomb = Y.Suppose an email contain the item or keyword {Bomb, Blast, Attack}. The itemsets of size one from this basket are {Bomb}, {Blast} and {Attack}; the itemsets of size two are {Bomb, Blast}, {Blast, Attack} and {Bomb, Attack}; and itemsets of size three are {Bomb, Blast, Attack}. Some itemsets will appear in many different emails. For instance, bomb and blast will be found in many email samples: the itemset {bomb, blast} occurs frequently. The support for an itemset is the number of transactions where that itemset appears. Association rules are constructed from itemsets. For example, the itemset {Tense=past, Attack=Y} occurs in many of the transactions where {Bomb=Y} appear then the following association rule can be written
{Tense=past, Attack=Y, Bomb=Y}
->{Class=informative}
This rule's confidence is the percentage of transactions containing {Tense=past, Attack=Y,} that also contain {Bomb=Y}. The support for the rule is the number of transactions that contain {Tense=past, Attack=Y} and {Bomb=Y}.An association rule can have many items in its antecedent (left hand side) and many items in its consequent (right hand side). The rule {Tense=past, Attack=y, Bomb=Y}-> {Class=informative} has antecedent {Tense=past, Attack=Y,Bomb=Y}and consequent {Class=Informative}.



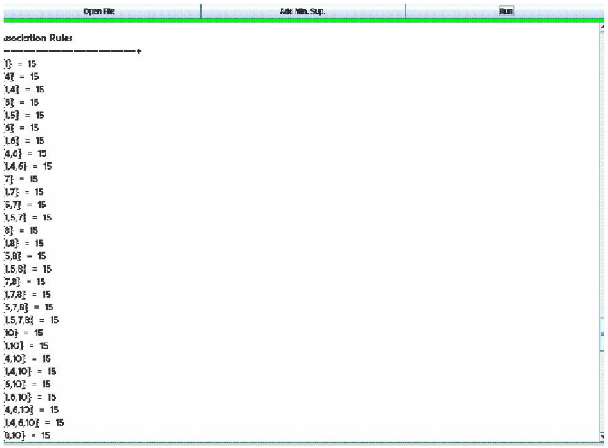
```
Reading input file
\\sample.txt
{1,4,6,8,10,13,15}
{1,5,7,8,10,12,15}
{2,4,6,8,10,13,16}
{3,5,6,9,10,12,16}
{1,5,7,9,11,13,14}
{2,4,6,9,11,13,16}
{1,5,7,9,11,12,15}
{1,4,6,9,10,12,15}
{3,5,6,8,10,12,16}
{3,4,7,9,11,12,15}
Number of records = 10
Number of columns = 16
```

**Fig.7. Reading Input file**



```
priori-T (Minimum support threshold = 20.0%)
--------------------------------------------
enerating K=1 large itemsets
enerating K=2 large itemsets
enerating K=3 large itemsets
enerating K=4 large itemsets
enerating K=5 large itemsets
enerating K=6 large itemsets

REQUENT (LARGE) ITEM SETS (with support count)
----------------------------------------------
1) {1} = 5
2) {2} = 2
3) {3} = 3
4) {4} = 5
5) {1,4} = 2
6) {2,4} = 2
7) {5} = 5
8) {1,5} = 3
9) {3,5} = 2
10) {6} = 6
11) {1,6} = 2
12) {2,6} = 2
13) {3,6} = 2
14) {4,6} = 4
15) {1,4,6} = 2
16) {2,4,6} = 2
17) {5,6} = 2
18) {3,5,6} = 2
19) {7} = 4
20) {1,7} = 3
21) {5,7} = 3
22) {1,5,7} = 3
```

**Fig.8. Frequent (Large) item sets generation**



```
ssociation Rules
-----------------+
1} = 15
4} = 15
1,4} = 15
5} = 15
1,5} = 15
6} = 15
1,6} = 15
4,6} = 15
1,4,6} = 15
7} = 15
1,7} = 15
5,7} = 15
1,5,7} = 15
8} = 15
1,8} = 15
5,8} = 15
1,5,8} = 15
7,8} = 15
1,7,8} = 15
5,7,8} = 15
1,5,7,8} = 15
10} = 15
1,10} = 15
4,10} = 15
1,4,10} = 15
6,10} = 15
1,6,10} = 15
4,6,10} = 15
1,4,6,10} = 15
8,10} = 15
```

**Fig.9.Apriori algorithm for Association rules generation**

The association rules generated are in numerical values hence the visualized output with respect to the output column of the preprocessing.

This Item sets are then used to generate Association Rules and one such rule is

**Tense=past, Attack=Y, Bomb =Y -> Email = Suspicious informative Email.**

**Tense=future, Attack= Y, Bomb =Y -> Email = Suspicious alert Email.**

**Tense=future, Attack= N, Bomb =N -> Email = Normal Email.**

These rules state that if there is Tense=past, attack &bomb then the email is suspected one with information that is it does not have any hazardous effect in future.

## 4. EXPERIMENTAL RESULTS

The application of data mining to the task of suspected email detection is done; experiments were carried out on a small email corpus. A mixture containing 1000 informative emails, 1000 alert emails and 1000 normal emails. The system was trained with the training dataset and the default support and confidence threshold were used. When the training process was finished, the top 20 best quality rules were taken as the final classification rules. Some of the rules generated by the Association rule based classification are:

**{Terrorist=y, Attack=y, Tense=past}→{Suspicious=informative}**

**{Threaten=y, Tense=future} →{Suspicious=alert}**

**Tense=future, Terrorist = Y, Bomb =Y → {Suspicious =alert}**

**{Tense=future, Attack=y, Blast=y} → {Suspicious=alert}**

**{Tense=past, Terrorist=y, Attack=y}→{Suspicious=informative}**

**{Tense=Past, Bomb=N, Terrorist=N, Blast=N→ Email=Normal}**

Output of Apriori algorithm will be a set of rules for each category. These rules were then used to classify the testing data. In the testing stage, rules generated in the training stage to be used to classify the incoming email. Extracted emails should be preprocessed before comparing with the generated rules. Processing such as lexical analysis, stoplist word removal and stemming should be applied to the extracted data. Resulting emails should be compared with each rule and the email is categorized to the most exactly matching category. If the left part of the rule matches with the email, count the category of the rule of the document. For this purpose, we will assign a priority value for each generated rule. Higher priority value will be assigned to large frequent itemsets and vice versa. Since we are assigning priority, we can increase the classification accuracy. If a rule matches with the email, corresponding priority value will be added to the category. Finally, the category with the highest value will be classification for that email.

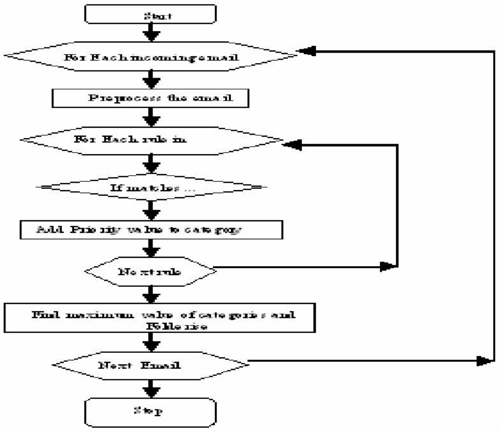
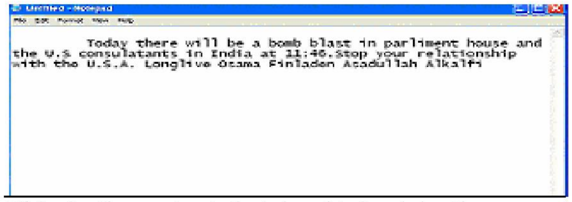


**Fig.10.Classifier Testing**

This is the input to the testing stage



```
          Today there will be a bomb blast in parliment house and
the U.S consulatants in India at 11:46.Stop your relationship
with the U.S.A. Longlive Osama Binladen Asadullah Alkalfi
```

This is the output that is obtained in the execution stage

The frequent itemset {Tense = future, Blast =Y, Bomb = Y} and the resulting association rule is

If Tense= future and Blast =Y and Bomb = Y then Email=Suspicious (alert). This is a Suspicious Email of alert type that is it will lead to any consequences in future.

## 5. CONCLUSION AND FUTURE WORKS

Email is an important vehicle for communication .It is one possible source of data from which potential problem can be detected. In this paper, we have employed Association rule mining based classification approach to detect deceptive communication in email text as informative or alert emails. We can find it that a simple Apriori algorithm can provide better classification result for suspicious email detection. In the near future, we plan to incorporate other techniques like different ways of feature selection, and Classification using other methods. One major advantage of the association rule based classifier is that it does not assume that terms are independent and its training is relatively fast. Furthermore, the rules are human understandable and easy to be maintained or pruned by human being. In this paper, a method of applying Association rule mining for suspected email detection is presented using keyword extraction and considering key attribute called Tense. The proposed work will be helpful for identifying the deceptive email and also assist the investigators to get the information in time to take effective actions to reduce the criminal activities.

A problem we faced when trying to test out new ideas dealing with email 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 email classification domain also suffers from the fact that it tries to reach general conclusion using very small data sets collected on a local scale.

## REFERENCES

**[ABSS00]** R.Agrawal, R. J. Bayardo, and R. Srikant. Athena, “Mining-based interactive management of text databases,” In Proc. 7th Int. Conf. Extending Database Technology (EDBT'00), pages 365-379, Konstanz, Germany, 2000.

**[AIS93]** R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” In *Proc.1993 ACM-SIGMOD Int. Conf. Management of Data*, pages 207–216, Washington, D.C., May 1993.

**[AS94]** R.Agrawal and R.Srikant, “Fast algorithms for mining association rules,” In Proc. 20th Int. Conf. Very Large Data Bases (VLDB'94), pages 487-499, Santiago, Chile, 1994.

**[BOO98]** G. Boone. “Concept features in re:agent, an intelligent email agent,” In Proc. 2nd Int. Conf. Autonomous Agents (Agents'98), pages 141-148, New York, 1998.

**[CDAR97]** S. Chakrabarti, B. E. Dom, R. Agrawal, and P. Raghavan, “Using taxonomy, discriminants, and signatures for navigating in text databases,” In Proc. 23rd Int. Conf. Very Large Data Bases, pages 446-455, Athens, GR, 1997.

**[CMSCT]**B.Cui, A.Mondal, J.Shen, G.Cong, and K.Tan, “On Effective Email Classification via Neural Networks,” In Proc. of DEXA, 2005, PP.85-94.

**[DLW00]** Y. Diao, H. Lu, and D. Wu, “A comparative study of classification-based personal e-mail filtering,” In Proc. 4th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD'00), pages 408-419, Kyoto, JP, 2000.

**[JHYZ05]** Jie Tang, Hang Li, Yunbo Cao and Zhaohui Tang, “Email Data Cleaning,” KDD’05, Chicago, USA.

**[JHMK]** Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques” Morgan Kaufmann Publishers”.

**[Joa98]** T. Joachims, “Text categorization with support vector machines: learning with many relevant features,” In Proc. 10th European Conf. Machine Learning (ECML'98), pages 137-142, Chemnitz, Germany, 1998.

**[KS05]** P.S.Keila and D.B.Skillicorn, “Detecting unusual and Deceptive Communication in Email,” Technical reports June, 2005.

**[LHM]** Liu,W. Hsu, and Y. Ma, “Integrating classification and association rule mining,” In *ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD’98)*, pages 80–86, New York City, NY, August 1998.

**[MOO2]** Maria-Luiza Antonie and Osmar R.Zaiane., “Text Document Categorization by Term Association,” IEEE International Conference on Data Mining (ICDM’2002), PP 19-26, Maebashi City, Japan, December 9-12, 2002.

**[OM02]** Osmar R.Zaiane ,Maria-Luiza Antonie,“Classifying Text documents by associating terms with text categories,” In Proceeding of the Thirteenth Australian Data base Conference (ADC’02), Melbourne, Australia , January 28-February 1, 2002.

**[Qui93]** J. R. Quinlan. C4. 5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.

**[Ren00]** J. Rennie, “An application of machine learning to e-mail filtering,” In Proc. KDD 2000 workshop on Text Mining, Boston, MA, 2000.

**[SDHH98]** M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. “A bayesian approach to filtering junk e-mail,” In Proc. AAAI'98 Workshop Learning for Text Categorization, Madison, Wisconsin, 1998.

**[SK00]**R. B. Segal and J. O. Kephart. Swiftfile, “ Intelligent assistant for organizing E-mail,” In AAAI 2000 Spring symposium on Adaptive User Interfaces, Stanford, CA, 2000.

**[WZH00]** K. Wang, S. Zhou, and Y. He, “Growing decision trees on support-less association rules,” In Proc. 6th ACM-SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'00), pages 265-269, Boston, MA, 2000.

**[WZL99]** K. Wang, S. Zhou, and S. C. Liew, “Building hierarchical classifiers using class proximity,” In Proc. 25th Int. Conf. Very Large Data Bases (VLDB'99), pages 363-374, Edinburgh, UK, 1999.

**[ZD]**Zan Huang and Daniel D.Zeng, “A Link Prediction Approach to Anomalous Email Detection,”