International Journal of Engineering & Technology, *7 (4) (2018) 2760-2765*

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.18486
Research paper

Spam classification by using association rule algorithm based on segmentation

Shahad Suhail Najam [1] *, Karim Hashim AL-Saedi [1]

[1] *Mustansiriyah University, College of Science, Department of Computer Science*
**Corresponding author E-mail:Shahad92.2015@gmail.com*
dr.karim@uomustansiriyah.edu.iq

Abstract

Email is a most widespread and active communication technique. The major purpose behind the success of email is the vast availability, facility of utilize, and affordability. Therefore this technology has be a susceptible to malicious attacks; Email is the most frequently applied delivery technique for malware. E-mail spam is one of the major problems of the Internet today, and get financial harm to companies and individual users is uncomfortable. Spam mail can be harmful as they may include malware & links to phishing Web sites. So necessary to divide spam from mail messages to a separate folder. In this paper utilize one of datamining mechanism is association rule algorithm.in association rule; pattern discovered based on relationship between item-sets. The dataset utilized in proposed system is Enron dataset is divided into two parts: spam and non-spam. For extract features from dataset used Term Frequency Invers Term Frequency (TFIDF) method. For reduce dimensionality of feature space use Information Gain (IG) method.

Keywords:*Spam; Association Rule; Information Gain; Term Frequency Invers Term Frequency.*

1. Introduction

Spam is unsolicited junk mail sent over the internet. Nowadays spam appears as a new threat on the internet at the email system [1]. Email spam has become a major problem on the Internet industry as a whole and individuals suffering the effects of this problem. In addition to the cost of additional e-mail recipients time management, spam emails leads to consumption of computing and network resources, on the other hand, refers to network performance to network security problems in other words it has a direct impact e-mail availability[2].People who send spam or unwanted online called spammer. These are sent by commercial advertisers who may offer our uncomfortable suspicious lifestyle, or encourage unwanted actions, and here the intention to make email the user spends money. There is another type of spammer who sends a large number of e-mail messages that by pass the server or the user mailbox, here the intention to harm email service so that users cannot receive e-mail. This is what is called a denial of service (DOS) [3]. For the reasons stated, the spam filter is one of the most important newly security systems that have been used recently. Spam filter has become a necessity and of great importance. In General, the spam filters focusing on classifying the emails and deletes spam or throws spam emails in spam folders [4]. This paper proposed approach to identify spam filtering email spam or non-spam based on the content of this message.in this proposed used one of data mining technique is association rule mining. Association rule mining find relationship between items and depended on two criteria is support and confidence. Also used association rule based on quantity algorithm instated of Apriori algorithms because Apriori algorithms depended on frequency without quantity [5, 6]. Section two displays some of related work which covers the problem of a spam. Section three presents a background on spam. Section four presents the methodology of the proposed system. The evaluation of results is discussed in section five. The last section presents the conclusion.

2. Related work

There are many kinds of literature on spam email, these are:

1) S. Divya and T. Kumaresan, (2014) [7], they were presented a spam classifier using machine learning algorithms including NB, SVM, and KNN was also proposed. The dataset used was spam assassin which contains 6000 emails 3776 of which for training and 2224 emails for testing. The numbers of features used was 100 features. In addition to the body of an email message, the classification based on other fields of the email such as the subject and the form. The performance evaluation recorded for the three classifiers was: (NB: Accuracy = 99.46), (SVM: Accuracy= 96.90), (KNN: Accuracy= 96.20). On the contrary of the previous studies, NB gave a satisfying performance among the other learning methods.
2) Wang Y. et al., (2015) [8], suggested "A new Document and Term frequency combined Feature Selection method (DTFS)" to improve the performance of classification of e-mails this method have several steps .firstly,an existing optimal document frequency based feature selection method (ODFFS) and a predetermined threshold are applied to select the most discriminative features.Secondly, an existing optimal term frequency based feature selection (OTFFS) method and another predetermined threshold are applied to select more discriminative features.Finally, ODFFS and OTFFS are combined to select the remaining features.and proposed global best harmony oriented harmony search (GBHS)to improve the convergence rate of parameter optimization, a metaheuristic method to search these optimal

predetermined thresholds. In this paper applied Naive Bayesian (NB) and fuzzy Support Vector Machine (FSVM) classifiers on six datasets, which are CSDMC2010, Enron-spam PU2, PU3, Trec2007 and Ling spam.

3) Thu Zarphyu,NyeinNyein, (2016) [9], this paper used a new feature subset selection algorithm based on conditional mutual information approach to selected the effective feature subset. This proposed comparing a new feature selection algorithm with the other feature selection algorithms for purpose evaluated anew feature selection algorithm through applied these algorithms on standard dataset UCI and weak (Waikato environment for knowledge analysis). In this paper used naïve Bayes and J48 classifier with different feature selection methods for the purpose of classification. In this paper it was concluded although some algorithm can reduce features more, while classification accuracy is not good. The proposed conditional mutual information based feature selection algorithm produces the effective and small features with higher classification accuracies in several different dataset.

4) Tianda yang, Kai Qian et al, (2016) [10], proposed comparison between using both association rule and naïve Bayes classifier algorithms and just using naïve Bayes classifier for purpose spam filtering. In this paper applied association rules and naïve Bayes on Enron-spam dataset. It proposed to use map reduce program to handle the amount of words. Map reduce approach is to use <key, value> pairs and the groups that will be received in the reduce function will be grouped by the key.

Map : <$K_{1,}$ V_1>→ list < K_2, V_2>

Reduce: < $K_{2,}$ list (V_2) >→ list (K_3, V3)

To enhancing implementation to combine naïve Bayesclassifier and A Priori algorithm, the purpose for the enhancing implementation is to improve ham precision rate.

3. Background

a) Spam Filter

The filter classification techniques are essentially classified for two parts:

1) Non – Machine Learning Technique

Non-Machine Learning means the programmer explicitly tells the computer what to do with the information it is given. This is traditional programming, and makes use of loops, objects, and functions [11]. Some of the non-machine learning algorithms for classification of data are listed.

Blacklisting filter: Is check list, which helps to reduce spam from the mail server IP address for e-mail and maintain blacklists (known as the blacklist (RBL, DNSBL)) Real-time Scan mail address consists of a set of rules that, if address matches email, mail server to return messages that have blank from field, which lists a lot of titles in "field" from the same source, which has too many numbers in email addresses (somewhat fake addresses to create poplar method). You can also return messages to match the language code in header [12-13].

White list filter: list, which includes all the addresses that users want to always receive mail. The user can add e-mail addresses or entire domains, or functional areas. An interesting option is the white management tool automatically and eliminates the need for administrators to manually input approved addresses on the white-list ensures that mail from certain senders or domains are never flagged as spam [12].

Signature based system filter: compares any incoming email to a known spam by computing its signature.

Mail Header Checking: In this method is a set of rules that is matched with the mail header to track whether spam or ham. If you match your system, then calls the server and route messages containing an empty field of “From”, confliction in “To”, confliction in “Subject” etc. [12-13].

2) Machine Learning Technique

Enables the computer to learn by itself without being programmed. Machine learning algorithms are more efficient in contrast to those of non-machine learning. Machine learning work in similar way like data mining, both acquire knowledge from data and find relevance in the data. Machine learning algorithms can be categorizedinto supervised and unsupervised algorithms. Some of the supervised machine learning algorithms for classification of data is listed below [14].

Clustering: is class from a technique used to separate groups of objects or relatively notes called groups. It classifies an object or control in this way so that objects in the same group as more similar to each other than those in the other group. Weak, and there are many who used the algorithms group aggregate method — that is, a hierarchical grouping method is used for distance (usually yoklidian) to determine the distance between each pair of object. A widely used means of grouping such as step by step to extract the advantage (or dictionary for learning) in the test phase many distance learning techniques. Performance depends on number of blocks due to incorrect input options, May lead to poor performance [12].

Incremental algorithm Decision tree algorithm (ID3): that is used to build a fixed set of notes citations (dataset) than to the tree used for incremental assessment test. And represents all note her features or attributes or class to which it belongs. The following decision tree representation decision tree leaf node contains the name of the category. A contract is a contract decision sheet. These include the contract condition involving attribute with sub with possible value for that attributeID3 used to measure gain information to be held. For information refer to the ability of certain attribute to separate categories of training examples. The upper part is gaining information, feature greater capacity for train control. Gets information using entropy as a measure to calculate the amount of uncertainty in the dataset . It builds a tree faster and attributes that are short enough to classify data. But they suffer from the problem of excessive if it was small [13].

Support Vector Machine Algorithm (SVM): is the machine under the supervision of technical learning that is used for classification and regression. In this land each data item as a point in n-dimensional space, where n = the number of features. In this method turns into higher dimensions of the original data and then looking for hyperplane optimization (frontier) which separates groups of one category from another clear enough gap as possible [16].

b) Classifiers in spam mail filtering

There are numerous types of Classifiers that are intended for the purpose of classifying spam e-mails or whispered this is mainly classified into two categories are essentially those being: content based Classifiers and non-content classifiers.

1) Content based Classifiers:-These classifiers also known as handmade spam works and these species that are classified in spams as carrying the content or information you store. It checks for the text in the body of the e-mail message, and then URL address. They also see the mail header like this thread to categorize text. It performs the task of classification of text using text processing in terms of removing HTML tags and calculates frequency Tokenizing and Word to determine the probability of word to find out if a particular mail is spam or not [16,17].

2) Non-Content based Classifiers: - In this type of classifiers that automatic filter is installed in this classification depends on the human recipient. Happening in this classification of the sender's name, address etc [17].

c) concepts about datamining

Data mining and knowledge discovery technology term, meaning make extracting useful information from raw data set. Data mining is part of knowledge discovery as in.

1) Collection of Raw Data: Dataset can collect data from various sources such as online and offline, social media sources, banks, and retail sector etc.

2) Data Selection: Select the relevant data that are used for the analysis.

3) Data Pre-Processing Data cleansing to remove any kind of noise, fake or missing value of data.
4) Transformation: Data turn out appropriately so that it can carry out the mining operation.
5) Data Mining: Extract relevant patterns of data using a different data mining techniques.
6) Evaluation: Extracted patterns are analyzed for the truthworthiness of the patterns and its relevance.
7) Knowledge: The procedure above mine relevant knowledge from raw data set. Different techniques can be knowledge representation [14-15].

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential pattern[14].

4. Methodology

The proposed system in this paper, titled "spam detection by using association rule based on segment". The objectives of using this system are to emails classification into spam email or non-spam email. This system consist of three modules; preprocessing modules, training modules, testing modules. Each of these modules contains several sub modules and components.

a) Preprocessing module

The preprocessing module contains three components. The first component is the tokenization, which holds the text email from dataset splitting the email into words. The second component is stop word removal, this component is designed to removal the stop word from emails such as (the, what, you etc.) Because these stop words not load any useful information. The third component is stemming, this component work to return each word in email into his root.

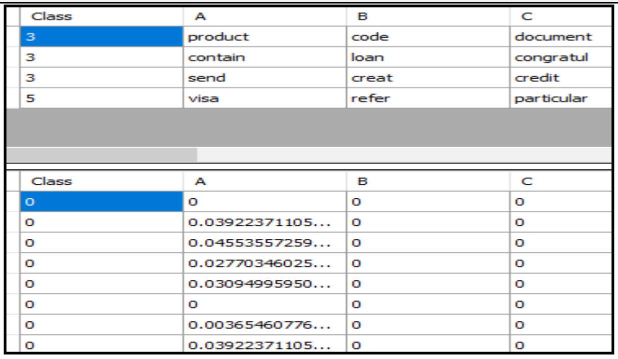
b) Training module

The main function of this module is to prepare the dataset which are obtained from previous modules to be easy classifying dataset in the testing module. In this module using three components, first component is Term Frequency Inverse Document Frequency (TFIDF) method for features extraction, second component is Information Gain (IG) for feature selection and third component association rule for classification.

1) TFIDF methods for feature extraction

This first component of training data set module. There is a need to this Feature extraction component to data representation because it is very hard to do calculations with text data. The representation should have to reveal the actual statistics for text data. Also a representation of the data in a way that even the actual statistics are converted to text data to value. Moreover, it should facilitate the classification and functions should be simple enough to implement. In this proposed systems used Term Frequency Invers Document Frequency (T_FID_F) for feature extraction to extract the distinctive features of spam and non-spam based on dataset. The first part T_F (t, d) is simply to calculate the number of items each word appears in each email. The second part ID_F(N, n)is the inverse ratio of emails containing the term t and total number of e-mail messages in the dataset. The implementation is illustrated in the algorithm (1) and show result in figure (1).

```
Algorithm (1) $T_FID_F$
Input
$(E_1, E_2, \dots, E_N)$ // All email from dataset
Output
F//Extraction features from all emails in dataset based on $(T_FID_F)$
Begin
Step1 :
For $E_i$ do// Compute $T_F$ for each email from dataset
    $T_F \leftarrow \frac{(T_w,e)}{(N_w,e)}$…$(T_w$,e) // Number of times the word appears in email
    $(N_w$,e) //Total number of words in the email
End for
Step2 :
for $E_i$ do// Compute$ID_F$ for each email from dataset
    $ID_F \leftarrow log\frac{N}{n_i}$ …. N // Number of emails
    $n_i$//Number of email that contain word
End for
Step3:
For $E_i$ do// Compute $T_FID_F$ for each email in dataset
    $T_FID_F \leftarrow T_F$* log $\frac{N}{n_i}$
End for
```



Class	A	B	C
3	product	code	document
3	contain	loan	congratul
3	send	creat	credit
5	visa	refer	particular

Class	A	B	C
0	0	0	0
0	0.03922371105...	0	0
0	0.04553557259...	0	0
0	0.02770346025...	0	0
0	0.03094995950...	0	0
0	0	0	0
0	0.00365460776...	0	0
0	0.03922371105...	0	0

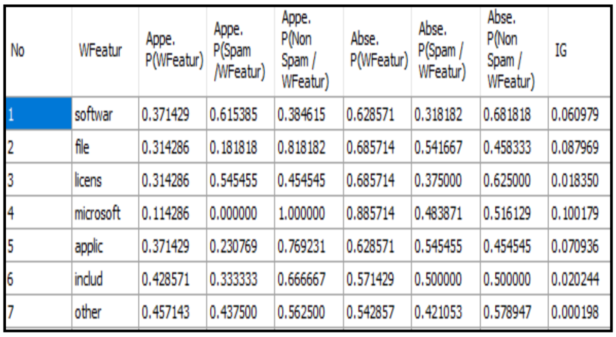
Fig. 1: Show Result From (T_FID_F).

2) IG algorithm for feature selection

Feature selection is second component of training module. Feature selection applies to taking only useful subset of features without changing their original forms. Feature selections have several techniques. In the proposed system used the information gain method, this method provided weight for each features, making the parser facility features multiple selection system based on its weight without limiting the number of random features, it allows the module to deal with many of the features in the same Time, and initiate a dynamic handling with multiple features. Algorithm (2) outlines the steps of calculating IG per each feature and figure (2) show result from IG.

```
Algorithm (2) feature selection by IG
Input
F// all features extraction from $T_FID_F$
Output
Features based on IG
BEGIN
Step1:
For $C_i$ do // Compute probabilities for each class(spam , non -spam)
    P(c) $\leftarrow \frac{frequency\ (c)}{N}$…Frequency (c) // number of class (spam, non-spam) email
    N // total of email
End for
Step2:
  a) For $F_i$ do //compute probabilities appearance each feature in all mail
    P (F) $\leftarrow \frac{f_i}{N}$ ….P(F) // probabilities for each features
    $f_i$ // appearance of features in all mails
    N// total of mails
  End for
  b) For Fi do //compute probabilities appearance each feature in email class
    P(c |F) $\leftarrow \frac{cf_i}{Nf_i}$ …. $cf_i$ // appearance of feature in each class (spam, non-spam) email
    $Nf_i$ // appearance of feature in all mails
  End for
Step3:
  a) For $F_i$do //compute probability absence for each features in all mails
    P$(F^-) \leftarrow \frac{f_i^-}{N}$… $f_i^-$ // absence offeatures in all mails
    N// total of emails
  End for
  b) For $F_i$ do //compute probability absence for each features in each class email
    P(C |$F^-) \leftarrow \frac{cf_i^-}{Nf_i^-}$ ….$cf_i^-$// absence of feature in each class mails
    $Nf_i^-$ // absence of feature in all mails
  End for
Step4:
For $F_i$ do //Compute entropy
  a) Total – Entropy $\leftarrow$ - p ( c) * $log_2$ p(c)
```

```
b)    Feature_ Entropy ←- p ( c) * $log_2$ p(f)
End for
Step5:
For $F_i$ do //Compute IG
IG←total Entropy – feature –Entropy
```



No	WFeatur	Appe. P(WFeatur)	Appe. P(Spam /WFeatur)	Appe. P(Non Spam / WFeatur)	Abse. P(WFeatur)	Abse. P(Spam / WFeatur)	Abse. P(Non Spam / WFeatur)	IG
1	softwar	0.371429	0.615385	0.384615	0.628571	0.318182	0.681818	0.060979
2	file	0.314286	0.181818	0.818182	0.685714	0.541667	0.458333	0.087969
3	licens	0.314286	0.545455	0.454545	0.685714	0.375000	0.625000	0.018350
4	microsoft	0.114286	0.000000	1.000000	0.885714	0.483871	0.516129	0.100179
5	applic	0.371429	0.230769	0.769231	0.628571	0.545455	0.454545	0.070936
6	includ	0.428571	0.333333	0.666667	0.571429	0.500000	0.500000	0.020244
7	other	0.457143	0.437500	0.562500	0.542857	0.421053	0.578947	0.000198

Fig. 2: Show Result from (IG).

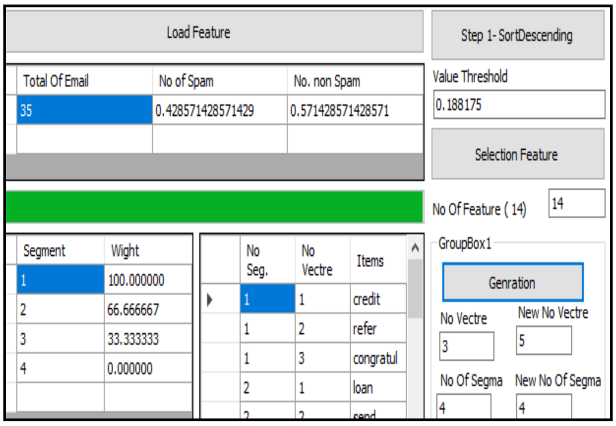
3) Implementation segments for features

After applying (IG) algorithm for all features, a number of segmentation of these features will work depended on IG because the IG algorithm depends on the select of high weights. Each segment consists of a specified number of features. If the number of features is larger than the number allocated in each segment, then if the number of remaining features is very small add these features in the last segment but the number of remaining feature large will add the new segment (this process operate according to certain threshold).Benefit from the process of segmentation in order to facilitate the work of the algorithm (AR) and obtain a high evaluation. The implementation is illustrated in the algorithm (3).

Algorithm (3) segmentation

```
Input:
IG features based on specific threshold ,
$N_w$ //number of features that apply the threshold value
$N_{collected}$ // the number of specified segment
$X_1$// the number of features in each segment
$X_0$// the number of remaining features after segment
Output:
$S_{IG //}$ segment based on IG features
Step1:
$X_1$←$N_w$/ $N_{collected}$
$X_0$←$N_w$ mod $N_{collected}$
If $X_0$ <> 0then
If $X_0$> 3 then
If ($N_{collected}$ * $X_1$+1) <$N_w$then
Add $seg_1$
Else
Add features in the least segment
Return $S_{IG}$
End.
```

In the segmentation phase it depends on threshold for information gain value (IG) and numbers of words for each segment and calculated weighted for each segment as shown in figure (4).



Total Of Email	No of Spam	No. non Spam
35	0.428571428571429	0.571428571428571

Value Threshold: 0.188175

No Of Feature (14): 14

Segment	Wight
1	100.000000
2	66.666667
3	33.333333
4	0.000000

No Seg.	No Vectre	Items
1	1	credit
1	2	refer
1	3	congratul
2	1	loan
2	2	send

No Vectre: 3

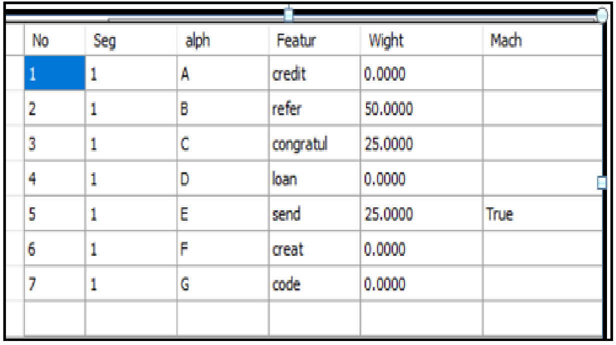
New No Vectre: 5

No Of Segma: 4

New No Of Segma: 4

Fig. 4: Offer Generation for Sample Segments and Weighted for Each Segment.

After that calculated weights for each feature in each segment show this in the figure (5).



No	Seg	alph	Featur	Wight	Mach
1	1	A	credit	0.0000	
2	1	B	refer	50.0000	
3	1	C	congratul	25.0000	
4	1	D	loan	0.0000	
5	1	E	send	25.0000	True
6	1	F	creat	0.0000	
7	1	G	code	0.0000	

Fig. 5: Weights for Sample Features of One Segment.

c)   Implementation for association rule

This proposal system applied association rule algorithm (AR) is belonging to data mining to purpose classification. The result of the algorithms predicts the incoming email into spam or non-spam. The next section explains about association rules algorithms.

1)   Association rules Generation based on quantity algorithm

In this proposal uses association rules to classify emails into spam or non-spam. Use association rules based on quantity algorithm this algorithms depended on weight and quantity to item added to frequency item in the dataset. Uses these algorithms instead of a prior algorithm because a prior algorithm depends on only frequency items regardless quantities and weighted this calculated cause weakness. Where strong association rules depended on both support and confidence. Steps association rules Generation based on quantity:

Step1: find quantity item sets in database

1) Transaction in database works for him scanned to calculated total quantity for each 1-itemset.
2) Calculated total quantity for each column in data set.
3) Calculated minimum quantity support (MQS) for each 1-itemset, L1 from through by using the following equation (1).

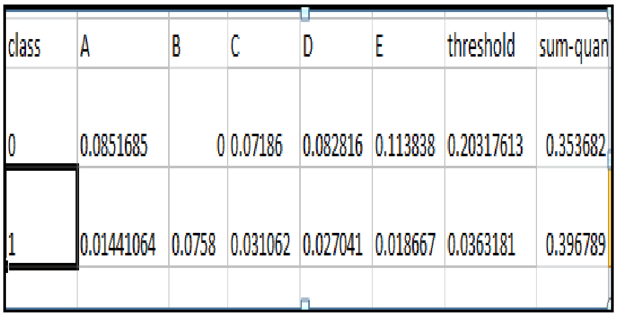
$$MQS=\frac{Quantity\ (1-itemset)}{totalquantityofitems} \qquad (1)$$

4) Selected threshold of quantities through less values for MQS
5) Discovery total frequency 2-itemsets, L2 algorithm uses (L1*L1) to generate (C2) candidates set of 2-itemsets.
6) The next, transaction in data base works for him scanned and totals quantity is calculated for each candidate item set inC2 and MQS calculated for each L2,2-itemset. It continues for the end rules.
7) The end calculated for each rulesweighted quantity confidence(WQC) even find strong association rules to satisfy

for each minimum quantity support and minimum quantity confidence this can be calculated from following equation(2). The implementation is illustrated in the algorithm (4) and figure (6) , figure (7) and figure(8) show this steps

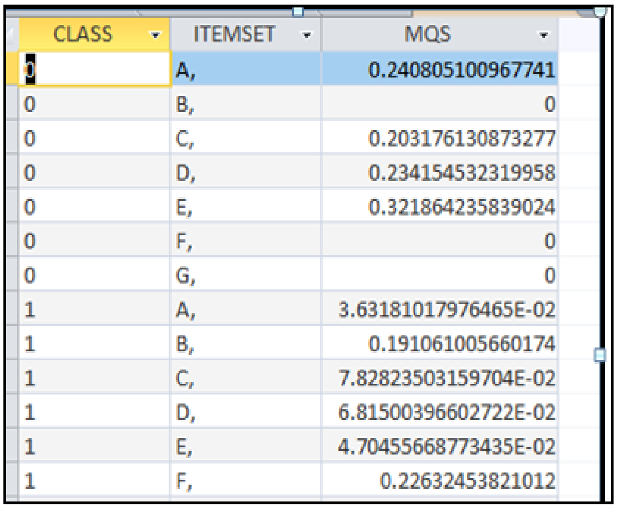
$$WQC=\frac{quantity(k-itemset)}{quantity((k-1)-itemset)}*100\% \qquad (2)$$

```
Algorithm (4) Quantity Based Association Rule Mining Algorithm
Input : value of feature , class (spam ,non-spam)
Output: rules
Begin
Step 1:
For $C_i$ do
    a)  Sum Total Quantity ($TQ_i$) for all feature
    b)  For Fi do
    •   MQS= $\frac{F}{TQ}$
End for
    •   Find minimum threshold from $S_i$
End for
Step 2:
    a)  Generate all rule based on number of feature =2
    b)  Sum Total Quantity ($TQ_i$ )
   for Ci do
    c)  Threshold for each class
    d)  For R where $R_i$>2 do
    1)  Len column= len of rule
    2)  End for
    3)  For Fi  value do
Find minimum for len two column to feature
Find total minimum based on above step
End for
    1)  Temp MQS=$\frac{totalmax}{TQ}$
    2)  If MQS > threshold $_i$ then take this rule and MQS
End for
Step3:
For $R_i$ do
Find confidence based on Temp MQS
End for
End
```



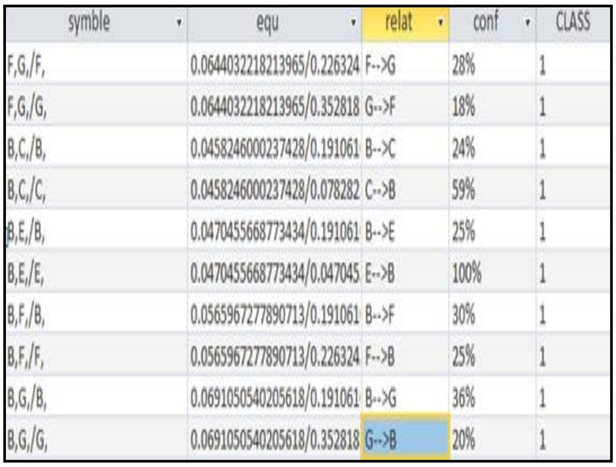
class	A	B	C	D	E	threshold	sum-quan
0	0.0851685	0	0.07186	0.082816	0.113838	0.20317613	0.353682
1	0.01441064	0.0758	0.031062	0.027041	0.018667	0.0363181	0.396789

Fig. 6: Calculated Total Quantity for Sample Features.



CLASS	ITEMSET	MQS
0	A,	0.240805100967741
0	B,	0
0	C,	0.203176130873277
0	D,	0.23415453231995‌8
0	E,	0.321864235839024
0	F,	0
0	G,	0
1	A,	3.63181017976465E-02
1	B,	0.191061005660174
1	C,	7.82823503159704E-02
1	D,	6.81500396602722E-02
1	E,	4.70455668773435E-02
1	F,	0.22632453821012

Fig. 7: Calculated MQS for Sample Feature.



symble	equ	relat	conf	CLASS
F,G,/F,	0.0644032218213965/0.226324	F-->G	28%	1
F,G,/G,	0.0644032218213965/0.352818	G-->F	18%	1
B,C,/B,	0.0458246000237428/0.191061	B-->C	24%	1
B,C,/C,	0.0458246000237428/0.078282	C-->B	59%	1
B,E,/B,	0.0470455668773434/0.191061	B-->E	25%	1
B,E,/E,	0.0470455668773434/0.047045	E-->B	100%	1
B,F,/B,	0.0565967277890713/0.191061	B-->F	30%	1
B,F,/F,	0.0565967277890713/0.226324	F-->B	25%	1
B,G,/B,	0.0691050540205618/0.191061	B-->G	36%	1
B,G,/G,	0.0691050540205618/0.352818	G-->B	20%	1

Fig. 8: Extracted Rules for Sample Feature.

2) Association rules based on quantity algorithms for segments All steps mentioned above apply and add new step .calculate the weighted for all the features in each segment. The benefit of calculation the weighted of each feature until it determines which segment is testing .the weight of each features is calculated within the segment by taking the feature and calculating its frequency with all the rules that were formed in the each segment. The implementation calculated weighted for all the features in each segment in the algorithm (5).

```
Algorithm (5) weighted algorithm
Input: features , rules
Output: weighted for each feature
Begin
Step1:
For Fi do
If Fi in rule then
$sum_{Li}$ + =1
if Fi in rule then
$sum_{Ri}$ + =1
End for
Step2:
For Fi do
$P_{Li}$= $sum_{Li}$/ no. rule
$P_{RJ}$ = $sum_{Ri}$ / no. rule
$W_j = \frac{Pli+PRj}{2}$
End for
End
```

3) Tasting for association rules quantity algorithms for segment

The first incoming emails words are taken after initial preprocessing the email words are match with the first segment. If the match occurs, all the rules in this segment are taken. And find max confidence for all rules. The implementation is illustrated in the algorithm (6).

```
Algorithm(6):Testing
Input: email, segment
Output: classification
Begin
Step1:
Read Email
Xfiled // preprocessing email
For each seg do
Get feature weight (call algorithm(3.9))
End for
Selected need weight for seg
For each rule in seg max do
If xfile in rule then
Xrule add (rule , count)
End for
Step2:
```

```
Find max content
, classify
End
```

5. Spam detection evaluation

The performance of classifier based on confusion matrix (TP, TN, FP, FN) experiments on testing sample present in the next section.

a) Accuracy: this is the ratio between the sum of true prediction was divided by the total emails. The below table show accuracy result for experiment.

Table 1: Accuracy Results on Testing Sample

Threshold value	No. features	No. segments	Association rule
0.198117	12	4	90%
0.188175	14	4	94%
0.209727	10	3	92%
0.154149	24	8	91%

b) Error rate : this is the ratio between the sum of false prediction was divided by the total Emails. The below table2 show accuracy result for experiment.

Table 2: Error Rate Results on Testing Sample

Threshold value	No. features	No. segments	Association rule
0.198117	12	4	91%
0.188175	14	4	90%
0.209727	10	3	92%
0.154149	24	8	93%

c) Precision: this is the ratio between the sum of true positive was divided by the total number of positive prediction. The below table3 show precision result for experiment.

Table 3: Precision Results on Testing Sample

Threshold Value	No. features	No. segments	Association rule
0.198117	12	4	92%
0.188175	14	4	94%
0.209727	10	3	91%
0.154149	24	8	93%

d) Recall: this is the ratio between the sum of true positive was divided by the total number of true positive and false negative prediction. The below table4 show precision result for experiment.

Table 4: Recall Results on Testing Sample

Threshold Value	No. features	No. segments	Association rule
0.198117	12	4	90%
0.188175	14	4	92%
0.209727	10	3	93%
0.154149	24	8	94%

6. Conclusion

After building email Spam Filter for detected the Spam and Ham email the following list are concluded: The implementation of association rules algorithms leads to the creation of a large number of item-set as well as requiring large space and less accuracy, so the dataset was segmented for a more accurate classification using the association rule algorithms this step give high accuracy for classification .And the implementation of the operation of the segment based on the information gain algorithm; because the work of information gain based on the selection of features that have high weights values. Classifiers performance is enhancing with a bigger training sample size.

Refrences

[1] AakankshaSharaff, Naresh Kumar Nagwani, And Kunal Swami," Impact Of Feature Selection Technique On Email Classification",June 2015

[2] Jon Kågström, “Improving Naive Bayesian Spam Filtering”, M.Sc. Thesis, Mid Sweden University Department For Information Technology And Media, Spring 2005

[3] Ciphertrust, “Spam: A Security Issue”, Ciphertrust, Inc. White Paper, December 2003.

[4] M. Basavaraju, And R. Prabhakar, “A Novel Method Of Spam Mail Detection Using Text Based Clustering Approach”, International Journal Of Computer Applications (0975 – 8887) Volume 5, No.4, August 2010.

[5] Jiawei Han, Michelinekamber and Jian Pei," Data Mining Concepts and Techniques ", third Edition, 2012.

[6] Karim Al-Saedi, S. Manickam, S. Ramadass, W. Al-Salihy and A. Almomani, 2013. "Research Proposal: An Intrusion Detection System Alert Reduction And Assessment Framework Based On Data Mining" Journal Of Computer Science. Volume 9, Issue 4. Pp: 421-426. New York, Usa.

[7] S. Divya, And T. Kumaresan, “Email Spam Classification Using Machine Learning Algorithm”, International Journal Of Innovative Research In Computer And Communication Engineering, Vol.2, Special Issue 1, March 2014.

[8] Wang Y., Liu Y., Feng L., and Zhu X., "Novel Feature Selection Method Based On Harmony Search for Email Classification", Knowledge-Based Systems 73, 311–32, 2015.https://doi.org/10.1016/j.knosys.2014.10.013.

[9] Thuzarphyu,Nyein,"Performance Comparison Of Feature Selection Methods",Yangon Technological University,2016.

[10] Tianda Yang, Kaiqian Et Al," Spam Filtering Using Association Rules And Naïve Bayes Classifier",Ieee,2016.

[11] G.Kaur, R. K. Gurm, “A Survey On Classification Techniques In Internet Environment”, In International Journal Of Advance Research In Computer And Communication Engineering, Vol. 5, No. 3, Pp. 589–593, 2016.

[12] V.Christina, S.Karpagvalli, G.Suganya,"Email Spam Filtering Using Supervised Machine Learning Techniques", 2010.

[13] Eshabansal,PradeepKumarbhai,"A Survey Of Various Machine Learning AlgorithmsOn Email Spamming ",Irf International Conference, 8th January, ,Dravidian University, 8th January, January 2017.

[14] Harjot Kaur*, Er. Prince Verma, "Survey on E-Mail Spam Detection Using Supervised Approach with Feature Selection", Nternational Journal of Engineering Sciences & Research Technology, April 2017.

[15] P. Verma And D. Kumar, “Association Rule Mining Algorithm’s Variant Analysis,” In International Journal Of Computer Applicaation, Vol. 78, No. 14, Pp. 26–34, 2013.

[16] SeongwookYoun, And Dennis Mcleod, "Spam Email ClassificationUsing An Adaptive Ontology", Journal Of Software, Vol. 2, No. 3, Pp.43-55, September 2007.