ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2012.0050



# AN EFFECTIVE SPAM FILTERING FOR DYNAMIC MAIL MANAGEMENT SYSTEM

**S. Arun Mozhi Selvi[1] and R.S. Rajesh[2]**
[1]*Department of Information Technology, Dr. Sivanthi Aditanar College of Engineering, India*
E-mail: heyaruna@gmail.com
[2]*Department of Computer Science and Engineering, Manonmaniam Sundaranar University, India*
E-mail: rs_rajesh1@yahoo.co.in

***Abstract***
***Spam is commonly defined as unsolicited email messages and the goal of spam categorization is to distinguish between spam and legitimate email messages. The economics of spam details that the spammer has to target several recipients with identical and similar email messages. As a result a dynamic knowledge sharing effective defense against a substantial fraction of spam has to be designed which can alternate the burdens of frequent training stand alone spam filter. A weighted email attribute based classification is proposed to mainly focus to encounter the issues in normal email system. These type of classification helps to formulate an effective utilization of our email system by combining the concepts of Bayesian Spam Filtering Algorithm, Iterative Dichotmiser 3(ID3) Algorithm and Bloom Filter. The details captured by the system are processed to track the original sender causing disturbances and prefer them to block further mails from them. We have tested the effectiveness of our scheme by collecting offline data from Yahoo mail & Gmail dumps. This proposal is implemented using .net and sample user-Id for knowledge base.***

***Keywords:***
***Spam, Bayesian, IMAP, ID3***

## 1. INTRODUCTION

In this modern society all are spending their most of the time with internet, the reason behind this is it provides a easy way of communication with the people where ever they are and also people find a way for buying and selling their product through internet to make money without wasting their time as much as. The main criterion for this is Providing Security. Especially the email system is suffered with degraded quality of service due to rampant spam and fraudulent emails**.** Thus in order to avoid these types of problem a system is needed to extract only the needful information for the user as per his/her requirement and preferences. By doing this most of the unwanted mails from the mail user agent can be filtered to our notice which will be a great use for the user while viewing their regular mails.

### 1.1 INTERNET

The Internet is a global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP) to serve billions of users worldwide. It is a network of networks that consists of millions of private, public, academic, business, and government networks, of local to global scope, that are linked by a broad array of electronic and optical networking technologies. The Internet carries a vast range of information resources and services, such as the inter-linked hypertext documents of the World Wide Web (WWW) and the infrastructure to support electronic mail. The Internet has enabled or accelerated new forms of human interactions through instant messaging, Internet forums, and social networking. Online shopping has boomed both for major retail outlets and small artisans and traders. Business-to-business and financial services on the Internet affect supply chains across entire industries.

### 1.2 EMAIL

Electronic mail, commonly called email or e-mail, is a method of exchanging digital messages across the Internet or other computer networks. Originally, email was transmitted directly from one user to another computer. This required both computers to be online at the same time, a la instant messaging. Today's email systems are based on a store-and-forward model. Email servers accept, forward, deliver and store messages. Users no longer need be online simultaneously and need only connect briefly, typically to an email server, for as long as it takes to send or receive messages. An email message consists of two components, the message *header*, and the message *body*, which is the email's content. The message header contains control information, including, minimally, an originator's email address and one or more recipient addresses and the body contains the message itself as unstructured text; sometimes containing a signature block at the end. This is exactly the same as the body of a regular letter. The header is separated from the body by a blank line.

### 1.3 HOW SPAM FILTERING SYSTEM WORKS

There is no one specific algorithm for statistically determining whether or not a given e-mail message is in fact a spam message. As discussed earlier, the most prominent approach to spam classification involves the implementation of the Bayesian chain rule, also known as Bayesian filtering.

### 1.4 MOTIVATION

The Existing system still confuses us in working with our mailbox. The major part of the page holds the unwanted newsletters and advertisement Though there are certain packages helpful to extract the needful information they are not up to the users full satisfaction and also act as a spyware which totally upset's the user. There exists a strong call to design high-performance email filtering systems. A careful analysis of spam shows that the requirements of an efficient filtering system include: (1) accuracy (2) self-evolving capability (3) high-performance which needs to be completed quickly especially in large email or messaging systems. We are motivated by the inadequate classification speed of current anti-spam systems. Data have shown that the classification speeds of current spam filters fall far behind the growth of messages handled by servers. Based on this a system has to be proposed for an efficient spam filtering.

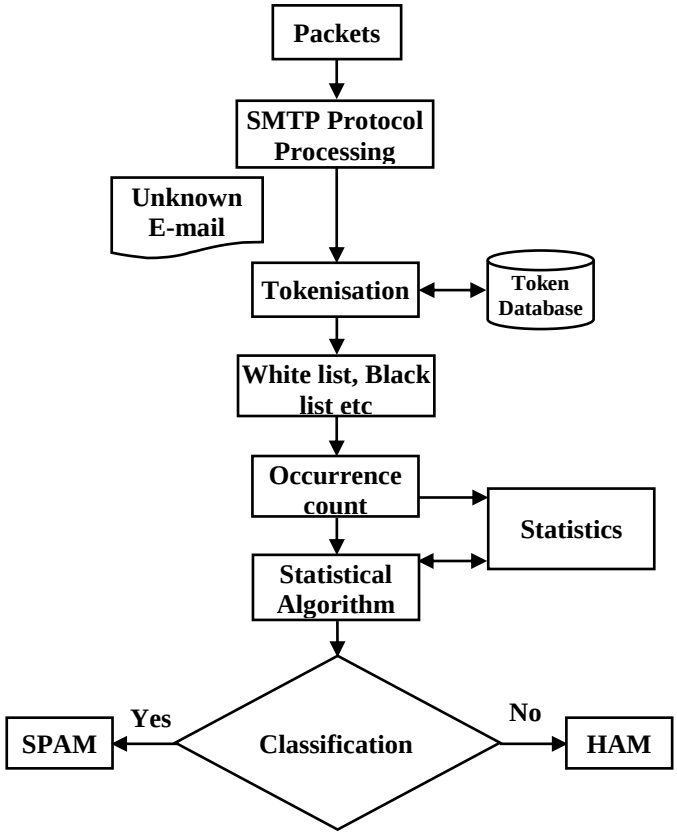


Fig.1. Work flow of the Spam Filtering System

### 1.5 PROBLEM STATEMENT

Most of the existing research focuses on the design of protocols, authentication methods; neural network based self-learning and statistical filtering. In contrast, we address the spam filtering issues from another perspective – improving the effectiveness by an efficient algorithm. They focus only towards the better improvement of acquiring the mail box information from spam mails. This system is mainly to overcome the difficulties faced by the current mail server agents. The system acts as an interface to the mail server and captures the mail information as per the user's requirement which in turn avoids advertisement, unwanted mails from reading and wasting the time working with large stuff of information dumped in mailbox. A weighted email attribute based classification is proposed to mainly focus to encounter the issues in normal email system. It makes the user to feel more securable by means of detecting and classifying such malicious mails when the user checks the inbox by notifying with different colors for spam (red), suspected (blue) and ham (green) mails**.** These type of classification helps to formulate an effective utilization of our email system.

### 1.6 OVERVIEW OF THE PAPER

The thesis is organized as follows section 2 the background and motivation of this research with the help of reference paper and internet. Section 3 introduces the proposed mechanism which describes the major work. Section 4 describes the experimental results.

## 2. BACKGROUND AND RELATED WORKS

By the inadequate classification speed of current anti-spam systems data have shown that the classification speeds of current spam filters fall far behind the growth of messages handled by servers. Based on this a system has to be proposed for an efficient spam filtering. From [1] the Decision tree data mining technique is chosen to classify the mails based on the any score or weight. From [2] Hash based lookup for the token in the scan list is chosen to improve the speed and efficiency. From [3] the basic spam filtering process for parsing the tokens of each mail in an effective manner. From [4] learnt to adapt the system under partial online supervision so that the efficiency may be improved on usage. From [5] a new concept of categorizing the mail into an unclassified category which is neither SPAM nor HAM.

Thus based on the survey made a system should act as an interface to the mail server and classifies mails as per the user's requirements. Mails, the user always want to read are placed under regular and those mails the user never wants to read are placed under spam. The unexpected mails that the user wants to get but which are not much important can be placed under suspected mails. Thus based on this classification can be done by an effective filtering mechanism by combining the concepts of Bayesian Spam Filtering Algorithm, Iterative Dichotmiser 3 Algorithm and Bloom Filter. Owing to this a system is created as a knowledge base for spam tokens which repeatedly occur in the spam mails. The probability of occurrence of such tokens are calculated using Bayesian algorithm and the output of it will be the input to the bloom filter which assigns weight for those tokens for the easy lookup in the knowledge base. Based on the above information and several attributes like From_Id, Subject, Body, To_Id, Sender's IP Address the mails are further classified into three categories as White_List (HAM), Gray_List(SUSPECTED) , Black_List (SPAM) with a help of id3 algorithm. These type of classification helps to formulate an effective utilization of our email system. This proposal is implemented using .net and sample user-Id for knowledge base.

## 3. PROPOSED MECHANISM

Based on the survey related to classification an effective spam filtering mechanism is proposed by combining the concepts of Bayesian Algorithm, Iterative Dichotmiser 3 Algorithm, weighted attribute algorithm and Bloom Filter. Owing to this a system is created as a knowledge base for spam tokens which repeatedly occur in the spam mails. The probability of occurrence of such tokens are calculated using Bayesian algorithm and the output of it will be the input to the bloom filter which assigns weight for those tokens for the easy lookup in the knowledge base. Based on the above information and several attributes like From_Id, Subject, Body, To_Id, Sender's IP Address the mails are further classified into three categories as White_List (HAM), Gray_List(SUSPECTED), Black_List (SPAM) with a help of ID3 algorithm. These type of classification helps to formulate an effective utilization of our email system. This proposal is implemented using .net and sample user-Id for knowledge base.

The mechanism flows through the following stages,

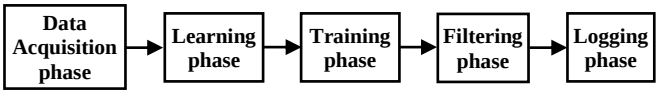


Fig.2. General System Model

## 3.1 DATA ACQUISITION PHASE

In this phase the no of mails of 4 different users are studied and the way they are categorized is captured. This Information is acquired from the Google and Yahoo dumps as they suffer a lot from the different types of spam. About 200 mails are analyzed and the mail Information retrieved from the current mail servers are extracted to and given to the next Learning Phase. Based on the Acquired data on different e-mail accounts, the following sample is shown in Table.1. From the subjects, it can be noted that some of the unwanted mails are under Ham mails (i.e. inbox). The analysis shows that 50% of the mails come under ham and the remaining 50% comes under spam. For example, "New SBI security update", "ICICI bank home loan" even though these mails are not much important they are under regular mail. Hence in order to reduce the amount of unwanted mails in inbox, an idea to classify the mails into a new category called suspected was decided. This category holds the mails that are not much important and they can be viewed separately at the user's convenience.

Table.1. Acquired information

| MAIL ID | SUBJECT | |
|---|---|---|
| | SPAM | HAM |
| zainabasiya@yahoo.co.uk | Free income<br>Re:<br>case<br>study:=?UTF | say hello to<br>extra income<br>this is so sad in<br>a way<br>Rest in peace |
| rizh2008@yahoo.in | Find your<br>perfect life<br>partner<br>Your 555USD<br>reward is here | A year of<br>innovation and<br>growth of<br>slideshare<br>Your six figure<br>balance |
| karthiga24@yahoo.com | Hey karthiga<br>VIAGRA<br>official | New SBI<br>security update<br>Just play to<br>collect |
| muthulakshmiit27@gmail.com | For you – 82%<br>OFF<br>Free!! | ICICI bank<br>home loan<br>Forward your<br>scan copy |

## 3.2 LEARNING PHASE

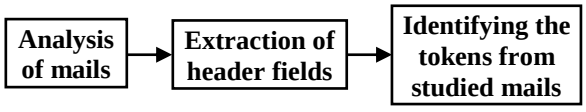


Fig.3. Flow Chart of the Learning Phase

Bayesian Spam filtering is a statistical technique of e-mail filtering. It makes use of naive base classifiers to identify spam e-mails. Bayesian classifiers work by correlating the use of tokens with spam and non-spam e-mails and then using Bayesian statistics to calculate the probability that an e-mail is spam or not. Rather than calculating the probability for all the tokens in the message. The list of spamminess tokens are identified by different users and evaluated for the scan list both for the subject and body of the message.

## 3.3 TRAINING PHASE

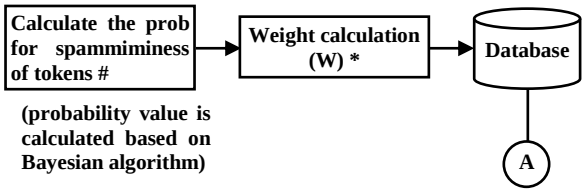


Fig.4. Flow Chart of the Training Phase

### *3.3.1 Probability Calculation for Spamminess Tokens Bayesian Theorem*:

To calculate the probability using Bayesian Theorem, first it needs to calculate the probability for individual words which is likely to be spam. This is calculated by using the following formula,

$$\Pr(S/W) = \frac{\Pr(W/S) * \Pr(S)}{\Pr(W/S) * \Pr(S) + \Pr(W/H) * \Pr(H)}$$

where,

- Pr(S/W) →probability that a message is spam knowing that word "x" is in it.
- Pr(S) →overall probability that any given message is spam
- Pr(W/S) →probability that the word "x" appears in spam messages
- Pr(H) →overall probability that any given message is not spam (i.e. ham)
- Pr(W/H) → probability that the word "x" appears in ham messages

### *3.3.2 Weight Calculation Based on Bloom Filter:*

In order to find the spamminess of the mail, a Bloom filter concept called weight methodology is introduced. The weight is calculated on the basis of the probability values calculated and the severity of the tokens that were analyzed during the learning phase. The weight methodology was obtained from the concept of bloom filter. In the Bloom filter, each tokens probability is considered to be associated with value 'w' for storing and retrieving, when used at the end to calculate a message's spamminess, a token's probability value 'w' is approximately mapped back to p. The value "w" represents the weight here.

The weight is calculated with a simple equation:

$$W = \text{Roundup}(P*100) \quad (1)$$

where,

- P = Probability of the token to be spam
- W = Weight assigned for the easy lookup

The following table shows the sample individual tokens of both subject fields, body their probability and weight for subjects.

Table.2 Acquired information Sample P and W value for the tokens found in both subject and body field

| Subject tokens | Probability (P) | Weight (W) | Body tokens | Probability (P) | Weight (W) |
|---|---|---|---|---|---|
| Buy | 0.525 | 5 | Password | 0.999 | 9 |
| Viagra | 0.999 | 9 | Fill in the info | 0.56 | 6 |
| Reply | 0.609 | 6 | Connect to | 0.7 | 7 |

The calculated weight is rated from 1 to 10 and the Threshold value is 5. The weight for each token is calculated in the Learning phase as per the severity of the token made in the analysis. The above values (token, probability, weight) both for subject and body are stored into database for further filtering.

## 3.4 FILTERING PHASE

The details learnt and calculated in the previous phase are given as the input to this filtering phase.

### *3.4.1 Filtering Algorithm (A Weighted Attribute Algorithm and ID3):*

The various header fields (critical attributes) and the message (usually body) are given as an input to the filtering algorithm – the algorithm used here to filter and classify the mails is Iterative Dichotmister3 (ID3). It is mathematical algorithm for building the decision tree. The tree should be built from the top to down, with no backtracking.

#### *3.4.1.1 A Weighted Attribute Algorithm (WAA):*

*WAA* states that

- If the message has the weight (w = 0), then it means Ham mail Notified by A
- If the message has the weight (w = 1 to 5), then it means Suspected mail Notified by B
- If the message has the weight (w > 5), then it means Spam mail Notified by C
- If any one condition is satisfied it exits the main algorithm.

#### *3.4.1.2 ID3 Algorithm:*

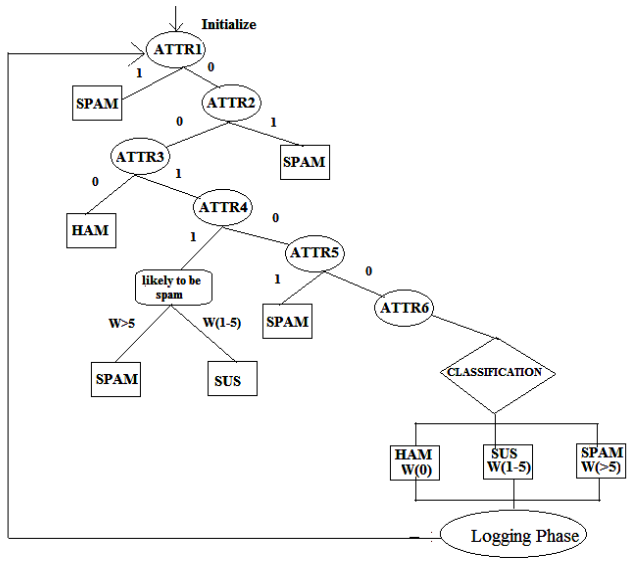


Fig.5. Flow Chart for ID3

**CRITICAL ATTRIBUTES:**
**Attribute 1** → Spam List
**1:** Spam id and Subject
**0:** Opposite situation
**Attribute 2** → "To id"
**1:** Not my id mark (mark.project11@gmail.com)
**0:** Opposite situation
**Attribute 3** → Contact List
**1:** From id not in Contact List
**0:** Opposite situation
**Attribute 4** → Subject contains Abnormal Keywords
**1:** Presence of Abnormal Keywords
**0:** Opposite situation
**Attribute 5** → Size of the Mail
**1:** No more than 6kB
**0:** Opposite situation
**Attribute 6** → Body checking
**CLASSIFICATION**
**TARGET ATTRIBUTES:**
⇨**HAM**
⇨**SPAM**
⇨**SUSPECTED**

Fig.6. List of Attributes for ID3 Algorithm

**Step 1:** Checks the List of Spam id and Subject if 1 classifies as C 0 step2

**Step 2:** Checks the TO id with user id if 1 step3 0 classifies as A

**Step 3:** Check with contact List if 1 classifies as A 0 step4

**Step 4:** Check with subject scan list if 1goto WAA 0 step5

**Step 5:** Check size <6kB 1classifies as C 0 step6

**Step 6:** Check with body scan list if 1 goto WAA 0 classifies as A

The cumulative weight is calculated in the algorithm when it reaches the Step 4 and Step 5 so that when the weight reaches the threshold the algorithm directly classifies rather than checking all the tokens, thus improves the efficiency.

## 3.5 LOGGING PHASE

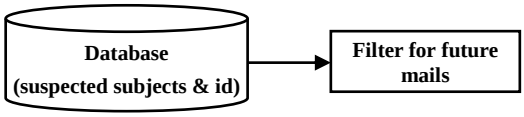


Fig.7. Flow Chart of the Logging Phase

This is the phase where all the details are logged in a file for the future use and maintenance.

### *3.5.1 Monitoring Database:*

The database table contains the suspected **id** and **subject**, which is stored from the mail that has already came to the inbox which is filtered out and then classified that it is spam. So, in future when the mails are coming from the same id are with the same subject is automatically redirected to the spam folder instead of checking the mails with all critical attributes and then finally redirects to the spam folder. This makes the filtering process much more efficient to the mail server.

## 4. IMPLEMENTATION

The proposed mechanism was implemented in .net platform and SQL server with the help of the 4 sample user's and their id. Based on the feedback of those sample users's the analysis is made based on these 4 categories.

- Current mail classification based on the no of mails
- Proposed Mail Classification based on the no of mails
- False positive Analysis for current mail server versus proposed system
- False Negative Analysis for current mail server versus proposed system
- Accuracy Analysis for current mail server versus proposed system

## 5. RESULT ANALYSIS

No of mail ids Analyzed: 5

- Id1 = heyaruna@gmail.com,
- Id2 = zainabasiya@yahoo.co.uk,
- Id3 = rizh2008@yahoo.in,
- Id4 = karthiga24@yahoo.in,
- Id5 = muthulakshmiit27@gmail.com

Analysis are made by the User's feedback for each mail id user

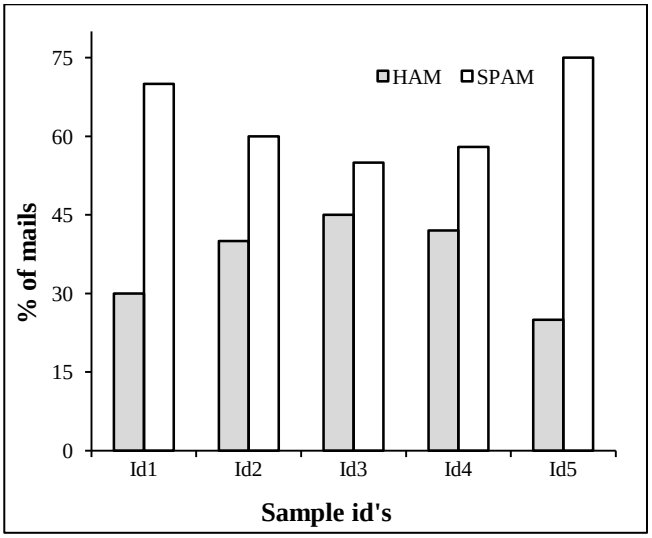


Fig.8. Current mail classification based on the no of mail

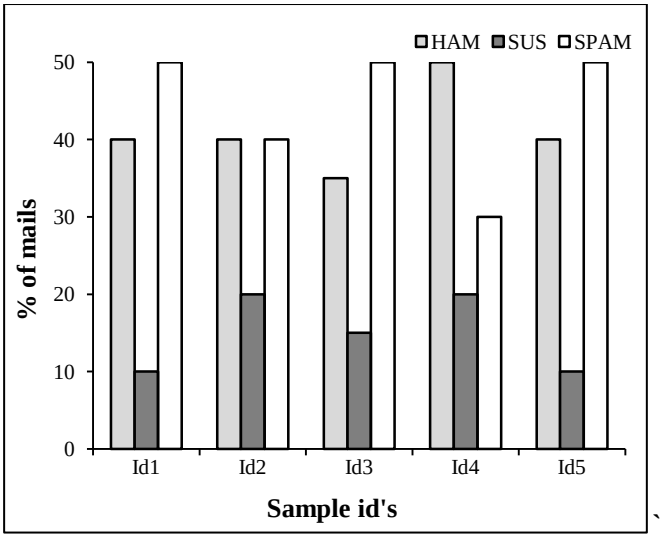


Fig.9. Proposed Mail Classification based on the no of Mails

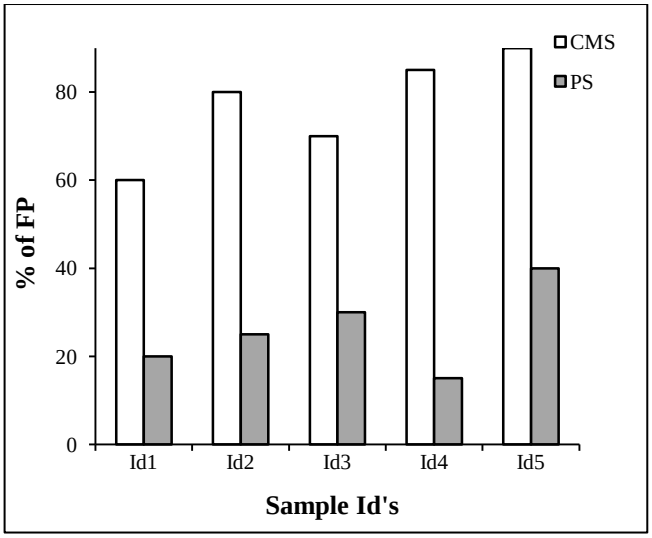


Fig.10. False positive Analysis for current mail server vs. proposed system

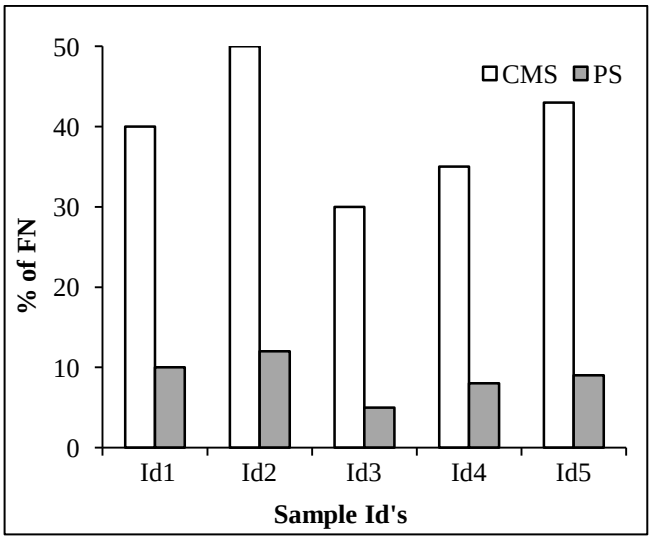


Fig.11. False Negative Analysis for current mail server vs. proposed system

False Positive (FP) – Classifying or identifying a ham mail as spam mail

False Negative (FN) – Classifying or identifying a spam mail as ham mail

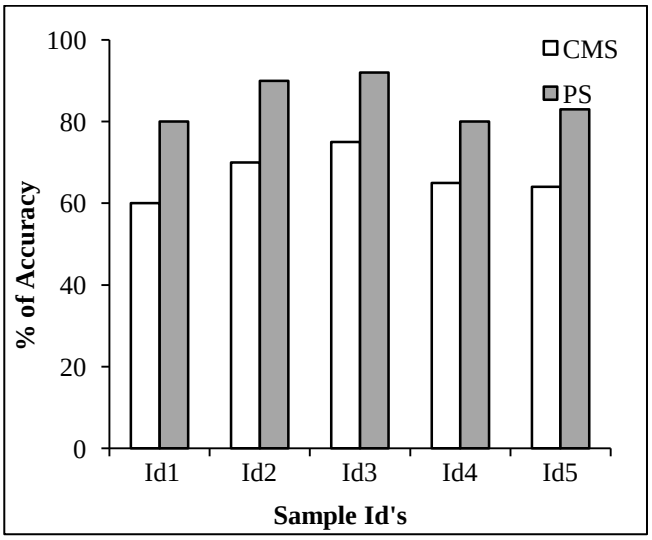


Fig.12. Accuracy Analysis for current mail server vs. proposed system

## 6. DISCUSSION

Thus from the above results it can be inferred that the % of false positives and false negatives in the current mail servers can be reduced. The Accuracy is also improved for large dataset. Thus it can be concluded that the major part of the inbox in current mail server is with spam messages which are considerably avoided in the proposed Algorithm.

## 7. CONCLUSION

This proposal is mainly to focus to encounter the issues in normal email system. These types of classification help to formulate an effective utilization of our current email system. Based on the implementation results the false positives and false negatives can be reduced gradually with the help of the logging phase in acquiring the original sender details. Based on this the Accuracy is also improved for large dataset.

## 8. FUTURE ENHANCEMENT

This proposal can be enhanced with more no of samples with more efficient Data Mining Technique. The implementation can be worked out in the mails servers for testing the effectiveness of the dynamic filtering system.

## REFERENCES

[1] Jhy-Jian Sheu "An Efficient Two-Phase Spam Filtering Method Based on E-mail Categorization", *International Journal of Network Security*, Vol. 9, No. 1, pp.34-43, 2009.

[2] Zhenyu Zhong and Kang Li "Speed Up Statistical Spam Filter by Approximation", *IEEE Transactions on Computers*, Vol. 60, No. 1, pp. 120 – 134, 2011.

[3] Yan Luo, "Workload Characterization of Spam Email Filtering System", *International Journal of Network Security and its Application*, Vol. 2, No. 1, pp. 22 – 41, 2010.

[4] Aris Kosmopoulos, Georgios Paliouras, Ion Androutsopoulos "Adaptive Spam Filtering Using Only Naïve Bayes Text Classifiers", *Spam Filtering Challenge Competition, Fifth Conference on Email and Anti-Spam*, Vol. 2, No.1, 2008.

[5] Brian whitworth and Tong Liu, "Channel E-mail: A Sociotechnical Response to Spam", *IEEE Computer Society*, Vol. 42, No. 7, pp. 63-71, 2009.

[6] Naresh Kumar Nagwani and Ashok Bhansali "An object oriented Email clustering model using weighted similarities between email attributes", *International Journal of research and reviews in Computer Science*, Vol. 1, No. 2, 2010.