# An Efficient E-Mail Generalization Scheme For Unsolicited Mail

A.Padma Priya 1 and Prof. R. Kalaivani2
1- M.Tech II (CSE), 2- Assistant Professor

**ABSTRACT:**

*Recently, the number of spam mails is exponentially growing. It affects the costs of organizations and annoying the e-mail recipient. Spammers always try to find the way to avoid filtering out from the email system. At the same time, as an email recipient or network system/administrator, we try to have an effective spam mail filtering technique to catch the spam mails. The problems of spam mail filtering are that each user has different perspective toward spam mails; so there are many types of spam mails, while the challenge is how to detect the various types and forms of spam mails. In this paper, the spam mail detected based on the subject of the spam mail. The information from the spam messages also can be used to filter spam mails and it can give higher accuracy than the keyword-based method does.*

***Key Points:*** *Spam detection, Filtering Technique, Subject Based Technique.*

## 1. Introduction

The application of internet grows rapidly in day to day life of every person. E-mail service commonly used for all types of peoples like students, business people, organizations and several peoples. The mails can be a normal mail and also a unsolicited mail. These unsolicited mails are sending by the spammers. The mail ID of the users can be collected from chat rooms, websites, or some other social web sites. They didn't send the mail with their own mail address. By using the fake id they send the mails. This creates the unnecessary traffic on the network. This creates the wastage of time to readers. It is a kind of advertisement mails. The first part of the mail is different from the normal mail. The central mail server can take part of the spam detection. But it is a cumulative e-mail. So, the bottleneck problem will occur and the performances also slow down. The one of the anti spam technique is filtering concept. In filtering concept black list and the white list are maintained. The black lists are the lists which are considered as spammers. If the message coming from the black list are automatically stored into the spam box.

White lists are the list which contains the trusted parties list of e-mail address. If the message coming from the white list is automatically accepted. The next technique is keyword based technique. The specific words, messages, images, are stored into the database. If the mail has those same phrases it would be identified. If suppose modern spammers can use different keywords that could not be identified. It is the main drawback of keyword based technique. Also the e-mail layout structure is based on the email abstraction. If the particular spam mail abstraction is not stored in the database and spammer try with some other abstraction then the spam mail is not able to identify. It is the main drawback in email abstraction scheme. The second section tells about the related works. The third section deals about system architecture and the next will deals about the design goals. The last section is conclusion.

## II. RELATED WORKS

Metsis.V, et al., has authored a paper about spam detection based on two methods. They are Support vector machine, Navie Bayes Classifier. The NBC and the SVM with different C parameters are trained on a set of 2000 emails with 1000 spam's and 1000 non-spam's and are tested on 200 new emails with 100 in each class.SVM separate feature vectors into the two classes by finding a hyper plane with maximal margin. The feature which is closest to the hyper plane is called SV. The NBC is the most effective approach for learning to classify text document. The draw back in this method is the value of the C parameter is large then it is hard to identify [14].

Issac B, Jap W.J, has authored a paper about Implementing spam detection using Bayesian and porter Stemmer Keyword stripping. The Porter Stemmer developed by Martin porter at the University of Cambridge in 1980. It is a process for removing the commoner morphological and in flexional endings from words in English. It is based on the idea that the suffixes in the English language are mostly made up of a combination of smaller and simpler suffixes. If a suffix rule matched to a word, then the conditions attached to that rule are tested on what would be the resulting stem. It works based on number of vowel characters, which are followed be a consonant character in the stem, must be greater than one. The drawback is it still improve with different multiple keyword weights [2].

Sousa.p, Machado.A, Rocha.M has authored a paperof spam filtering technique by using keyword based technique. The keywords are stored in the database. Local filters are used to identify these words and

IJERT

segregate the spam mail. The main drawback is if the keyword is not in the data base it cannot be identified [16].

Chi-Yao Tseng, Pin-Chieh Sung, and Ming-Syan Chen has authored a paper about detecting the spam based on the email abstraction. It is to generate the e-mail abstraction using HTML content in e-mail, and this newly devised abstraction can more effectively capture the near-duplicate phenomenon of spams. Moreover, we design a complete spam detection system, which possesses an efficient near-duplicate matching scheme and a progressive update scheme. The progressive update scheme enables system Cosdes to keep the most up-to-date information for near-duplicate detection. If the spammer send a spam mail with different email abstraction which is not stored in the database then the spam is not able to identify.

## III. DESIGN GOALS

Our design goal contains two main methods. The first method is the e-mail generalization and the second is to identify the spam mail based on the subject.

### A) Email Generalization

In Email generalization, email abstraction i.e., the html content of email is extracted from the mail. This method is called structure abstraction generation. SAG is composed of three major phases, Tag Extraction Phase, Tag Reordering Phase, and Appending Phase. In Tag Extraction Phase, the name of each HTML tag is extracted, and tag attributes and attribute values are eliminated. In addition, each paragraph of text without any tag embedded is transformed and then inserted into Anchor Set, and the first 1,023 valid tags are concatenated to form the tentative e-mail abstraction. Only the first 1,023 tags as the tag sequence are retained. The ordering of the tag sequence of an e-mail abstraction in Tag Reordering Phase. The main objective of appending tags is to reduce the probability that a ham is successfully matched with reported spam's when the tag length of an e-mail abstraction is short. Preprocessing is applied after the structure abstraction generation. The main objective of this preprocessing step is to remove tags that are common but not discriminative between e-mails. The other objective is to prevent malicious tag insertion attack, and thus the robustness of the proposed abstraction scheme can be further enhanced. The following sequence of operations is performed in the preprocessing step.

1. Front and rear tags are excluded.
2. Nonempty tags2 that have no corresponding start tags or end tags are deleted.Besides, mismatched nonempty tags are also deleted.
3. All empty tags2 are regarded as the same and are replaced by the newly created empty tag. Moreover, successive empty tags are pruned and only one empty tag is retained.
4. The pairs of nonempty tags enclosing nothing are removed.

### B) Design of Sp Table and Sp Tree

SpTable and SpTrees (sp stands for spam) are proposed to store large amounts of the e-mail abstractions of reported spams. Several SpTrees are the kernel of the database, and the e-mail abstractions of collected spams are maintained in the corresponding SpTrees. SpTree is designed to take charge of e-mail abstractions within a range of tag lengths. SpTable is created to record overall information of SpTrees

### c) Spam Detection System

There are three types of e-mails, reported spam, testing e-mail, and misclassified ham, required to be dealt with by Cosdes. When receiving a reported spam, Insertion Handler adds the e-mail abstraction of this spam into the database except that the reputation score of this reporter is too low. Whenever a new testing e-mail arrives, Matching Handler performs the near-duplicate detection with collected spams to do the judgment. Meanwhile, if a testing email is classified as a spam, this e-mail will be viewed as a reported spam and be added into the database. Moreover, Error Report Handler copes with feedback misclassified hams and adjusts Cosdes by degrading the reputation of related reporters to prevent malicious attacks. The main functionalities of deleting outdated spams are not only to alleviate the overhead of the server, but to reduce the risk of accidental deletion of hams. If the spammer send a spam mail with different email abstraction which is not stored in the database then the spam is not able to identify. To overcome this drawback, new method of detecting the spam mail is introduced. It is based on the subject of the email.

### D) Subject Based Spam Detection

The Subject of email should be suggestive enough of the contents of the article to enable a reader to make a decision whether to read the article based on the subject alone. In a typical long list of emails in an inbox, the subject line is the most prominent field, and so a meaningful subject is the most useful pieces of information we can include with our email to make it convenient to process by the recipient and deliver the message that we wish to convey.

Here is a list of some very popular words found in unsolicited email that can be used as spam-blocking words. Put the words below into our email filter list, to substantially reduce unwanted email. a) Lucky, b) Money, c) Misc, d) Health

This Fig: 1 shows the overall system architecture. The incoming mail is sent into two modules. It first sent into the layout generalization module. This extracts the layout of the e-mail by using

IJERT

the structure abstraction generation module. This generates the mail and sent to the preprocessing unit. This unit is to remove the tags that are common but discriminative e-mail. Then it is sent to the design of Sptable unit. This is created to record overall information.

The progressive update scheme enables system Cosdes to keep the most up-to-date information for near-duplicate detection. If the spammer sends a spam mail with different email abstraction which is not stored in the database then the spam is not able to identify. To overcome the disadvantage we combine the both generalization and subject based detection to detect the spam mails.

This will be sent to the spam detection module. This checks the information with the database. If the information is available then it is reported as spam mail. If new spam is identified it automatically updated in to database.

The statistical approach of the spam mail based on the subject is not usually the first one to try when they write spam filters. Most hackers' first instinct is to try to write software that recognizes individual properties of spam. This technique will catch 79.7% of the emails in spam corpus, with only 1.2% false positives.

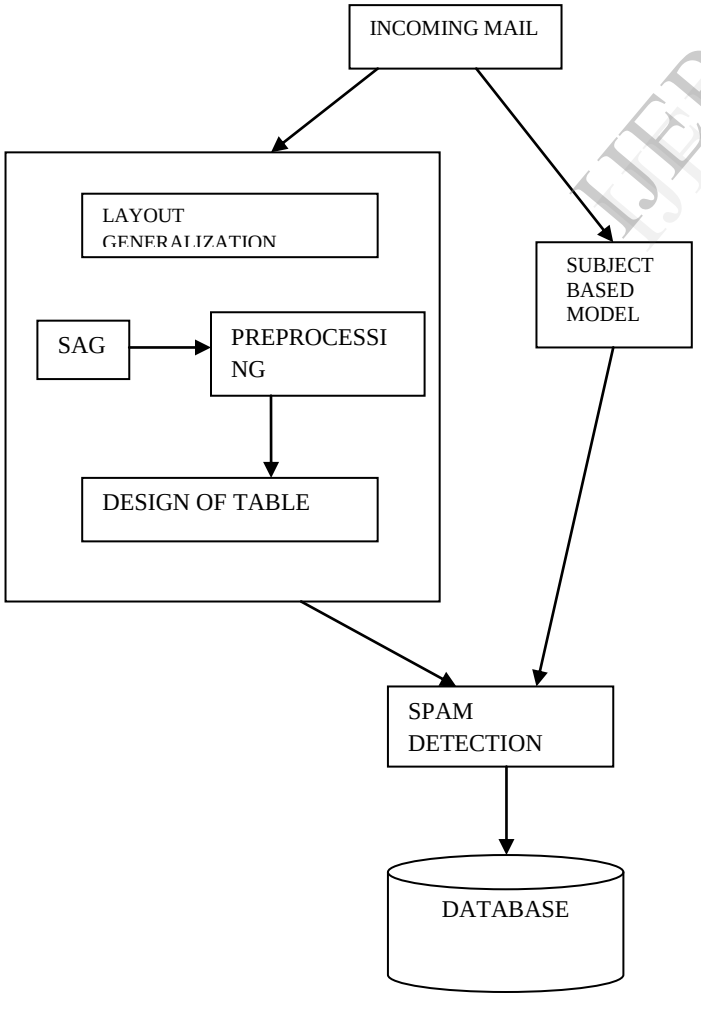


FIGURE 1: SYSTEM ARCHITECTURE

## IV. EXPERIMENTAL RESULTS

### Accuracy Evaluation

Cosdes reports 96.47 percent TP rate and 0.46 percent FP rate on average, which has the most outstanding performance of spam detection when compared to multidigest, digest and density. The TP rate of Digest is extremely high but the FP rate is unacceptable. In order to accelerate the process of near-duplicate matching, only a 32- byte code is used in Digest to represent each e-mail. The size of spam database is large, the 32-byte code is not discriminative to clearly distinguish each e-mail, and thus hams are easily mismatched with known spams. The multiple digests to represent each e-mail can be more robust against increased obfuscation effort by spammers, the FP rate of MultiDigest is even worse than that of Digest as the size of spam database is large. This is owing to the reason that MultiDigest separates each e-mail into a set of short strings.

As long as one digest in the huge spam database is similar to one of digests in the testing e-mail, this e-mail will be classified as a spam. In addition, the effectiveness of Digest and MultiDigest has not been validated by real e-mail streams

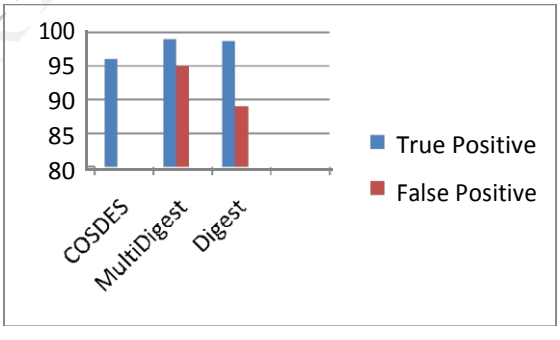


FIGURE 2 : THE COMPARISON OF AVERAGE DETECTION RESULTS OF COSDES WITH MULTIDIGEST AND DIGEST

Cosdes, which extract more essential information to represent each e-mail, and the newly devised e-mail abstraction can more effectively capture the near-duplicate phenomenon of spams with an acceptable FP rate. Cosdes, which evaluate the detection performance when either the sequence preprocessing step or the anchor-appending step of procedure SAG is removed. It can be observed in Fig. 2 that the performance almost does not degrade as it excludes the sequence preprocessing step. This consequence reveals that the proposed abstraction scheme has not been countered.

IJERT

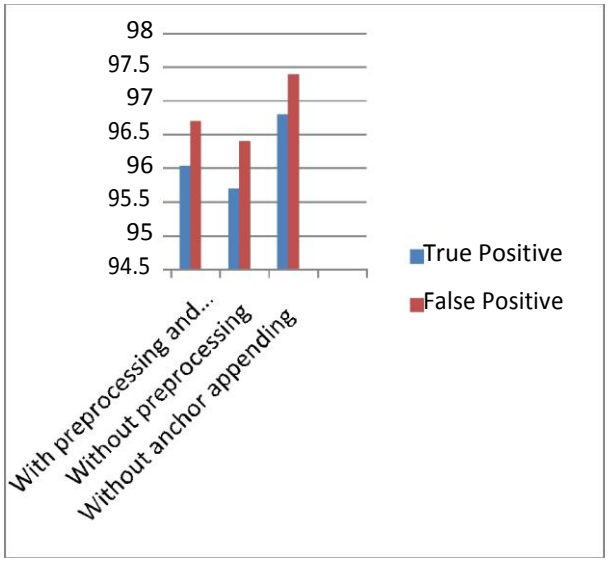


FIGURE 3: THE COMPARISON OF DETECTION PERFORMANCE IN IMPACT SEQUENCE PREPROCESSING AND ANCHOR-APPENDING

It can be seen in the Fig. 3 that the TP rate increases slightly but the FP rate is twice higher than that of the original situation as this process is removed. This is because there are several hams containing only a URL that normal users want to share with their friends. If the anchor-appending process is removed, these e-mails will be misclassified as spams, and thereby the FP rate deteriorates.

To overcome this disadvantage, the spam mail is detected based on the subject of the mail. The information from the spam messages also can be used to filter spam mails and it can give higher accuracy than the cosdes based spam detection. Based on the subject of the mail, the spam mail is detected based on the subject by which the spammer sends the mail. If the spammer sends a spam mail with different email abstraction which is not stored in the database then the spam mail is not able to identify. The subject based spam detection will reduce the risk of accidental deletion of hams. So it provides the accuracy of 97.03% of detecting spam.

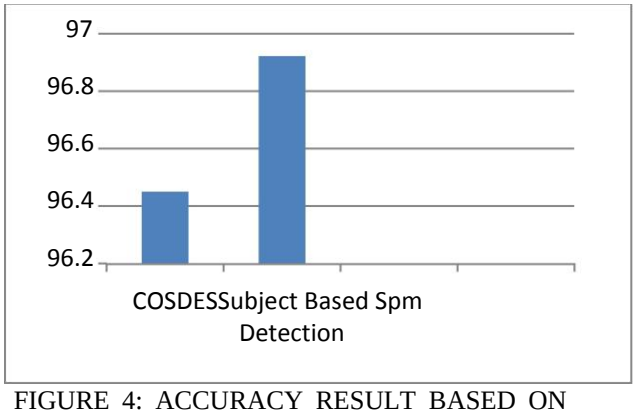


FIGURE 4: ACCURACY RESULT BASED ON COSDES AND SUBJECT BASED SPAM DETECTION

## V. CONCLUSION

In the field of collaborative spam filtering by near-duplicate detection, a superior e-mail abstraction scheme is required to more certainly catch the evolving nature of spam's. Compared to the existing methods in prior research, in this paper, it explores a more sophisticated and robust e-mail abstraction scheme, which considers e-mail layout structure to represent e-mails. The specific procedure SAG is proposed to generate the e-mail abstraction using HTML content in e-mail, and this newly-devised abstraction can more effectively capture the near-duplicate phenomenon of spam's. Moreover, a complete spam detection system Cosdes has been designed to efficiently process the near-duplicate matching and to progressively update the known spam database. Consequently, the most up-to-date information can be invariably kept to block subsequent near-duplicate spams. In the experimental results, it shows that Cosdes significantly outperforms competitive approaches, which indicates the feasibility of Cosdes in real-world applications. If the spammer sends a spam mail with different email abstraction which is not stored in the database then the spam is not able to identify. So, the information from the spam messages also can be used to filter spam mails and it can give higher accuracy.

## VI. REFERENCES

[1] Blanzieri.E et al.,(2008). A survey of learning-based techniques of email spam filtering. Artificial Intelligence Review, vol 29, No 1:pp.63-92.

[2] Biju Issac et al.,(2009), Implementing Spam Detection using Bayesian and Porter Stemmer Keyword Stripping Approaches, in Fourth International Conference on spam mai;,pp 124-130.

[3] Dat Tran et al.,(2007), Possibility Theory-Based Approach to Spam Email Detection, IEEE International Conference on Granular Computing,vol. 4,no.2,pp. 150-159.

[4] D. Sculley and G.M. Wachman, "Relaxed Online SVMs for Spam Filtering," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 415-422, 2007

[5] Fawcett,T.(2004), "In vivo spam filtering: A challenge problem for KDD, SIGKDD Explorations, vol. 5, no. 2, pp. 140-148.

[6] Gray.A and Haahr,M.(2004) Personalized, Collaborative Spam Filtering, in 1st Conf. on Email and Anti-Spam CEAS.

[7] Grossman.R, et al., (2002), Data Mining Standards Initiatives, Communications of ACM, vol. 45, no. 8, pp. 59-61.

[8] Hsin-Changn Yang,et al.,(2011), Post-Level Spam Detection for Social Bookmarking Web Sites, International Conference on Advances in Social Networks Analysis and Mining, ACM Press,pp. 20-29

[9] Kanich.C et al.,(2008):An Empirical Analysis of Spam Marketing Conversion, in Computer and Communications Security Conference (CCS08). ACM, pp. 27-31.

[10] Kaminsky,P.(2008), Defending against Sybil attacks via social networks, IEEE/ACM Trans. on

IJERT

Networking (TON), vol. 16, no. 3, pp.576-589.

[11] Kim.H et al.,(2005) Preventing session table explosion in packet inspection computers, IEEE Trans. on Computers, vol 17,no.5 pp. 238- 240.

[12] K. Yoshida, F. Adachi, T. Washio, H. Motoda, T. Homma, A. Nakashima, H. Fujikawa, and K. Yamazaki, "Density-Based Spam Detector," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 486-493, 2004

[13] Mendez,.J.et al.,(2008), A Comparative Impact Study of Attribute Selection Techniques on Naive Bayes Spam Filters, in 8th Industrial Conference on Data Mining, LNAI 5077, pp. 213-227.

[14] Metsis.V, et al.,(2006), Spam Filtering with Naive Bayes Which Naive Bayes? in Third Conf. on Email and Anti-Spam (CEAS), pp. 125-134.

[15] Nelson.B et al., (2008), Exploiting Machine Learning to Subvert Your Spam Filter, in 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats. ACM Press, pp. 1-9.

[16] Pedro Sousa, Artur Machado, Miguel Rocha, Paulo Cortez, Miguel Rio, "A Collaborative Approach for Spam Detection", Second International Conference on Evolving Internet,2010.

[17] Ryota Matsumoto et al.,(2004), Some Empirical Results on Two Spam Detection Methods, vol.6,no.1,pp. 153-160.

[18] Srisanyalak.B and O.Sornil(2007), An Artificial Immunity Spam Detection System, in 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats. ACM Press, pp. 1-9.

[19] T.R. Lynam and G.V. Cormack, "On-Line Spam Filter Fusion," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 123-130, 2006

[20] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam Filtering with Naive Bayes—Which Naive Bayes?" Proc. Third Conf. Email and Anti-Spam (CEAS), 2006.

[21] Yan Gao et al.,(2010), A Comprehensive Approach to Image Spam Detection: From Server to Client Solution, IEEE Transactions on Information Forensics and security, vol. 5, no. 4.

[22] Zhong.Z et al.(2008), ALPACAS:A Largescale Privacy-Aware Collaborative Antispam System, Proc. INFOCOM, pp. 556-564.

IJERT