International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol. 3, Issue. 3, May - June 2013 pp-1825-1827 ISSN: 2249-6645

A Novel Method for Collaborative Spam E- Mails

Sayyed Saissen[1], Sayeed Yasin[2]
[1]*M.Tech, Nimra College of Engineering & Technology, Vijayawada, A.P., India*
[2]*Asst.Professor, Dept.of CSE, Nimra College of Engineering & Technology, Vijayawada, A.P., India*

***ABSTRACT*:** With the increasing popularity of electronic mail (E-mail), several people and companies found it an easy way to distribute a massive amount of unsolicited messages to a tremendous number of users at a very low cost. These unwanted bulk messages or junk emails are called spam E- mails. The majority of spam mails that has been reported recently are unsolicited commercials promoting services and products including sexual enhancers, cheap drugs and herbal supplements, health insurance, travel tickets, hotel reservations, and software products. Spam E- mails has become an epidemic problem that can negatively affect the usability of electronic mail as a communication means. Besides wasting users time and effort to scan and delete these spam messages received; it consumes network bandwidth and storage space, slows down e-mail servers, and provides a medium to distribute harmful and/or offensive content. This paper presents a novel method for collaborative spam messages.

***KEYWORDS*:** COSDES, E- Mail, HTML, Spam.

I. INTRODUCTION

Internet is the most widely used area all over the world. In internet most widely used are Electronic mails (E-mails). E-mails play a major role for the communication between the users .The people who are using E- mails cannot verify the duplicate and near duplicate web documents creating the more problems on the web search engines. These documents will increase the space required to store the index, slow down the search results and the annoy users. According to the data availability on the internet, the huge data are shorts texts such that mobile phone short messages (SMS), instant messages, chat log, BBS titles etc.

The statistical information is given by the Information Industry Ministry of china that more than one billion mobile phone short messages are sent each day in Mainland China. You already know how much of email is spam, but here are a bunch of other factoids as per [1] you may not be aware of:

- Ninety percent of spam is in English. A year ago it was ninety six, so spam is getting more "international."
- Eighty eight percent of all spam is sent from bot nets (networks of compromised PCs).
- Ninety one of spam contains some form of link.
- Unsolicited newsletters are increasing and are now the second most common spam type.
- Spam from webmail services like Gmail and Hotmail is not as common as you might think. Only **0.7%** of spam is sent from the webmail accounts.
- One in 284 emails contain malware.
- One in 445 emails are phishing emails.
- As many as ninety five billion phishing emails were in circulation in 2010.
- Unfortunately, the status of duplicate and the near duplicate messages is very complex. Among these especially the near duplicates and spam mails.

These differences may result from various causes, such as:

1. Same contents appearing on various different sites are all crawled, processed and indexed.
2. Mistake introduced while parsing these noisy and loosely structured and noisy text (HTML page may contain ads., and it is known as shorting of semantics useful for parsing).
3. Manual typos (all information on Internet are created by people originally) and manual revising while being referred as well as reused.
4. Explicit modification to make the short messages suitable for difference usage.

Checking may be applicable manually when the scale of the repository is small. e.g. hundreds or hundreds or thousands of instances. When the amount of instances increases to millions and more, obviously, it becomes impossible for the human beings to check them one by one, which is tedious, costly and prone to error. Resorting to computers for such kind of repeatable job is desired, of which the core is an algorithm to measure the difference between any pair of the short messages, including duplicated and near duplicated ones.

II. RELATED WORK

Spam, or unwanted commercial email, has become an increasing problem in the recent years. Estimates suggest that perhaps seventy percent of all email traffic is spam. As spam clutters inboxes, time and effort must be devoted to either deleting it after it is received, or preventing it from even reaching the users [2]. In [3], the authors presented the results of combining classifier outputs for improving both accuracy and reducing false positives for the problem of spam detection. In [4], the authors specified that a set of independently developed spam filters may be combined in simple ways to provide substantially better filtering than any of the individual filters. In [5], the authors introduced a novel hybrid model, Partitioned Logistic Regression, which has several advantages over both naive Bayes and logistic regression. This model separates the original feature space into various disjoint feature groups.

In [6], the authors proposed a decentralized privacy preserving approach to spam filtering. The solution exploits robust digests to identify the messages that are a slight variation of one another and a structured peer-to-peer architecture between mail servers to collaboratively share knowledge about spam. In [7], the authors specified that the algorithm aims to detect spam web pages. In this algorithm, the web page gains the spam rank value through forward links, which are the links of reverse direction used in the traditional link-based algorithm. In [8], the authors have analyzed the TF-IDF algorithm to first find the relevancy of the comments with respect to the subject and further checking for repetition of the words in the comments. In [9], the authors have worked with Bayesian algorithm to filter e-mail spams. The major goal is to propose a model for an incremental spam filtering and test the model using the three training schemes.

The problem of unsolicited bulk email or spam is today well-known to every user of the Internet. Spam messages not only causes misuse of time and computational resources, thus leading to financial losses, but it is also often used to advertise illegal goods and services or to promote online frauds. To evaluate the performance of the highest probability SVM nearest neighbor classifier, this is an improvement over the SVM nearest neighbor classifier, on the task of the spam filtering. SVM nearest neighbor (SVM-NN) is the combination of the SVM and k-NN classifiers. Highest probability SVM nearest neighbor (HP-SVM-NN) classifier applied to the task of the spam filtering with variable relative error cost. The major strengths of the SVM are the training is relatively easy. No local optimal, unlike in the neural networks. It scales relatively well to high dimensional data and the tradeoff between the classifier complexity and error can be controlled explicitly.

III. PROPOSED WORK

In the proposed work, a complete collaborative spam detection system (COSDES) is introduced. The algorithm model of Cosdes is illustrated in Figure 1. Initially, three parameters- Tm (the maximum time span for reported spams being retained in the system), Td (the time span for triggering Deletion Handler), and Sth (the score threshold for determining spams) should be given for Cosdes. Before starting to do the spam detection, Cosdes collects feedback spams for the time Tm in advance to construct an initial database. Three major modules are included in Cosdes:

- Abstraction Generation Module
- Database Maintenance Module
- Spam Detection Module

With regard to the Abstraction Generation Module, each e-mail is converted to an e-mail abstraction by Structure Abstraction Generator with procedure SAG. Three types of action handlers- Deletion Handler, Insertion Handler, and Error Report Handler; are involved in Database Maintenance Module. Note that although the term "database" is used, the collection of the reported spams can be essentially stored in main memory to facilitate the process of matching. In addition, Matching Handler in the Spam Detection Module takes charge of determining results.

```
System Cosdes
Input: Tm: the maximum time span for reported spams being retained in
           the system,
       Td: the time span for triggering Deletion Handler,
       Sth: the score threshold for determining spams
1   switch (circumstance)
2   case: when receiving a reported spam
3     if (EA.reporter.SR > Sinitial);
4       Trigger Insertion Handler(EA);
5       Increase SR of the reporter in RepTable; // Rep: Reputation
6     break;
7   case: when receiving a testing email
8     Trigger Matching Handler(EA, Sth);
9     if (the testing email is classified as a spam);
10      Trigger Insertion Handler(EA);
11    break;
12  case: when receiving a misclassified ham
13    Trigger Error Report Handler(EA);
14    break;
15  case: for every Td
16    Trigger Deletion Handler(Tm);
17    break;
End
```

Figure 1: Algorithm for COSDES

Cosdes deals with four circumstances by the handlers, and the detailed procedure flow will be explained as follows: For Insertion Handler, initially, the corresponding SpTree is found in the SpTable according to the tag length of the inserted spam, and nowNode is assigned as the root of this SpTree. In lines 3 to 8, we iteratively insert the subsequences of the e-mail abstraction along the path from root to leaf. If nowNode is an internal node, the subsequence with 2i tags is inserted into the level i. Meanwhile, the hash value of this subsequence is then computed. Then, "nowNode" is assigned as the corresponding child node based on the type of the next tag. If the next tag is a start (or end) tag, nowNode is assigned as the left (or right) child node. Finally, when nowNode is processed to a leaf node, the subsequence with the remaining tags is stored.

The principal concept of collaborative spam detection is to collect the human judgment to block subsequent near-duplicate spam. To ensure the truthfulness of the spam reports and to prevent malicious attacks, we propose the reputation mechanism to evaluate the credit of each reporter. The fundamental idea of the reputation mechanism is to utilize the reputation table to maintain a reputation score "SR' of each reporter according to the previous reliability record. Each inserted spam is given a suspicion score equal to "SR" of the reporter. In such a context, when doing near-duplicate detection, if the sum of the suspicion scores of matched spams exceeds a predefined threshold, the testing e-mail will be classified as a spam.

IV. CONCLUSION

Collaborative Spam Detection System (COSDES) possesses an efficient near duplicate matching scheme and a progressive update scheme. The progressive update scheme not only adds in the new reported spams, but also removes obsolete ones in the database. With Cosdes maintaining an up-to-date spam database, the detection result of each incoming E-mail can be determined by the near duplicate similarity matching process. In addition, to withstand the intentional attacks, a reputation mechanism is also provided in Cosdes to ensure the truthfulness of user feedback.

REFERENCES

[1]. http://royal.pingdom.com/2011/01/19/ email-spamstatistics/
[2]. Message Labs Spam Intercepts data, 2006, http://www.messagelabs.com/publishedcontent/publish/threat_watch_d otcom_ en / threat_statistics/spam_intercepts/DA_114633.chp. html
[3]. Gray and M. Haahr, "Personalised, Collaborative Spam Filtering," Proc. First Conf. Email and Anti-Spam (CEAS), 2004.
[4]. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos, "An experimental comparison of Naive Bayesian and keyword-based anti-spam Filtering with personal e-mail messages", In Proc. Of SIGIR-2000, ACM, 2000.
[5]. M.-T. Chang, W.-T. Yih, and C. Meek, "Partitioned Logistic Regression for Spam Filtering," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data mining (KDD), pp. 97-105, 2008.
[6]. E. Damiani, S.D.C. di Vimercati, S. Paraboschi, and P. Samarati, "P2PBased Collaborative Spam Detection and Filtering," Proc. Fourth IEEE Int'l Conf. Peer-to-Peer Computing, pp. 176-183, 2004.
[7]. Chenmin Liang, Liyun Ru, Xiaoyan Zhu. R-SpamRank: "A Spam Detection Algorithm Based on Link Analysis", State Key Laboratory of Intelligent Technology and Systems (LITS), Department of Computer Science and Technology Department, Tsinghua University, Beijing, 100084, China
[8]. H. Drucker, D. Wu, and V.N. Vapnik, "Support Vector Machines for Spam Categorization," Proc. IEEE Trans. Neural Networks, pp. 1048- 1054, 1999.
[9]. Lixin Fu and Geetha Gali, "Classification Algorithm for Filtering Email spam", Recent Progress in Data Engineering and Internet Technology, 2012, Volume 157, 149-154, DOI: 10.1007/978-3-642- 28798-5_21