A Statistical Spam Filtering Scheme Based on Grid Platform[*]

Peng Liu[1], Jian-she Dong[2], and Wei Zhao[3]

[1] Grid Research Center, PLA University of Science and Technology,
Nanjing 21007, China
[2] School of Computer and Communication, Lanzhou University of
Technology, Lanzhou 730050, China
[3] Institute of Command Automation, PLA University of Science and
Technology, Nanjing 210007, China
liupeng00@mail.tsinghua.edu.org

Abstract. Spam is in spate, which accounts for over 60 percent of all emails in the world recently. Researchers are trying to develop ways to fight it but few are effective. The paper put forward a new filtering scheme based on grid technology and statistical method, which regards the user computers and email servers as nodes of the grid. They contribute and consume statistics information on the grid platform. If the number of copies of an email is obviously err from normal value, to flag it as a spam then can be a reasonable operation. As more and more nodes join the platform, the filtering precision can be further improved, just as the simulation study shows.

Keywords: spam, grid, CNGrid, statistics.

1 Introduction

Since the Internet and its application acquire rapid development, a series of problems related to Internet have accompanied, some of which may lead to a lot of troubles, for example, the Spam. The flooding of Spam will result in a mass of network resources being wasted, and the normal email corresponding being affected.

The problem of spam email is apparent to any frequent email user: unwanted, unsolicited bulk messages are emailed to a large number of users indiscriminately, which is similar to bulk mails sending the traditional postal service. In September 2001, 8% of all emails in US were spam. By July 2002, this fraction had increased to 35% [1]. More recent studies report that, in North America, a business user received 10 spam emails on average per day in 2003, and that this number is expected to grow by a factor of four by 2008 [2]. Furthermore, AOL and MSN report a daily blocking of 2.4 billion spam emails from reaching their customers' inboxes. This traffic corresponds to about 80% of daily incoming emails at AOL [3]. It is reported by the Anti-spam center of ISC [**4**] that in China a user received 19.33 spam emails on average

[*] This work was supported by the National Science Foundation of China under the grant No. 60403043.

O. Castillo et al. (Eds.): Theor. Adv. and Appl. of Fuzzy Logic, ASC 42, pp. 527–534, 2007.
springerlink.com

per week and 63.97% of all emails were spam in Mar. 2006, this is 2.03 spam emails more than Oct. 2005.

Spammers conduct marketing, commercial, and even unethical activities by sending out a huge amount of spam. This high volume is required as it is the only way to receive enough economical benefit. There is therefore a heavy maldistribution on e-mail traffic, making document space density a good index to identify spam. Although ordinary users seldom send more than 1000 similar e-mails, spammers have to send the same spam far more than that. Note that some of the unethical spam mail are said to be difficult to judge even for a human. However, the existence of over thousand identical e-mails makes the fact clear. Actually, experimental results reported in Section 4 showed that simple threshold is enough to distinguish spam from other e-mails.

The evident difference between spam mail and normal mail is that the same spam mail will be delivered to a large number of users, but most of normal only have one single receiver. Based on this observation, this paper presents a counting method based on CNGrid for spam mail filtering, In order to avoid normal mail being classified as spam mail, we also use a white list (WL) to improve the precise of spam mail filter.

The rest of the paper is organized as follows: In Section 2, we present the related works of spam mail filtering. In Section 3, we present the proposed filtering system in detail. Then, in Section 4, we show the results of the experiment which reveals the effectiveness of the proposed anti-spam filter. Finally, section 5 presents conclusions.

2 Related Works

Over the past few years, different approaches have been presented to provide resistance against spammers. Some of them use a Bayesian-like approach, or a rule-based approach, and some use a cryptographic solution to protect against spamming problem.

The simplest and most intuitive of all technique used to curb spam was to keep a blacklist of addresses to be blocked, or a white list of addresses to be allowed are also used. However this technique is not proved to be successful – since the spammers started sending spam mails either without the senders address or by spoofing the sender address.

Somebody suggests the method which to increase spammer's cost, such as filters and fight back (FFB) [5], slow senders method and penny per mail method [6]. The working of FFB resembles DoS (Denial of Services) a little bit. It sends junk messages to the spammers to increase their working load. The slow senders method and penny per mail method require all email sender to carry out a calculation which will consume their work time or to pay a little fee for each email. These methods must be support by new protocol, so they are not easy to be popularized.

Another one kind method is to distinguish the email sender's identity. Such as the ePrivacy E-mail open the standard [7] and Questions-Answering filtering. The ePrivacy request all senders to declare taking part in "the no sending spam" alliance. Every declarer will gain the figure signature and insert the signature into each email header to insure the sender's identity. An email without the signature will be chucked. The Questions-Answering method requires that the mail sender to fill in a table at a

Web page, otherwise the email with not be granted to send out. These methods are effective, but increased burden for senders and are easy to lose the legal mail.

Compared with the above method, the filtering methods are recipient by more people. Cohen [8] suggesting that incoming mail can be categorized according to its contents based on automatic learning rules. Some sophisticated rule based methods have good performance in spam filtering. SpamAssassin [9] is a successful case to filter spam emails based on rules. But the higher false positive ratio is its greatest shortcoming for

The concept of Bayesian Junk Mail filters suggested by Sahami et al. [10] got popularity. The filter was based on naive Bayes classifier. This method achieved a relative high degree of precision, but the recall was slightly low. It means that study have been found that more spam mails were classified as normal incorrectly. It was also found that outright deletions of spam brought about relatively high costs.

The cryptographic solution to protect against spamming problem was presented by Ioannidis [11]. In that solution the email address was encoded with certain policies. There policies were encrypted using symmetric keys and generated the message authentication code. The drawback to this solution was quite lengthy mail address, which proved to be difficult to adopt in commercial solutions. Not all methods presented for spam classification are suitable for both desk top based and server-based mail classification. The spam classification at the desktop is often more customizable and accurate, but such solutions often need too much computing and analysis and they are not suitable for massive spam mail process. The server-based mail classification should consider more about performance and avoiding normal mail being classified as spam mail.

3 Architecture of the Anti-spam Grid Base on CNGrid

In this paper we chose Grid [12] technique as the basis of the anti-spam system. It based on the following considerations: (1) Spam is delivered globally, so we need a global infrastructure to gather information on spam. (2) As central control system may result in bottlenecks, a collaboration of distributed services will efficiently serve local users. (3) It is a most dynamic environment that all the servers, clients, and e-mails keep changing all the time, so we need to form a virtual organization which is adaptive to changes.

The architecture of Anti-Spam Grid with distributed statistic of figure signature and distributed Bayesian filter. We call this architecture as Anti-Spam Grid. The whole system is built of Grid server, Anti-Spam Grid mail server (ASG Server) and Anti-Spam Grid client (ASG Client).

The Grid Server is with responsibility for the ASG Servers' scheme, registration and detecting the ASG Servers' running state. It is also with responsibility for the client user's registration, granting safe certification to client users and assigning ASG Server for client users which will serve it.

The ASG Server storages the fingerprint corpus with the designated range and provides index server schemed by the Grid Server. At the same time, it provides the searching service of approximate email and sharing of Bayesian knowledge repository.

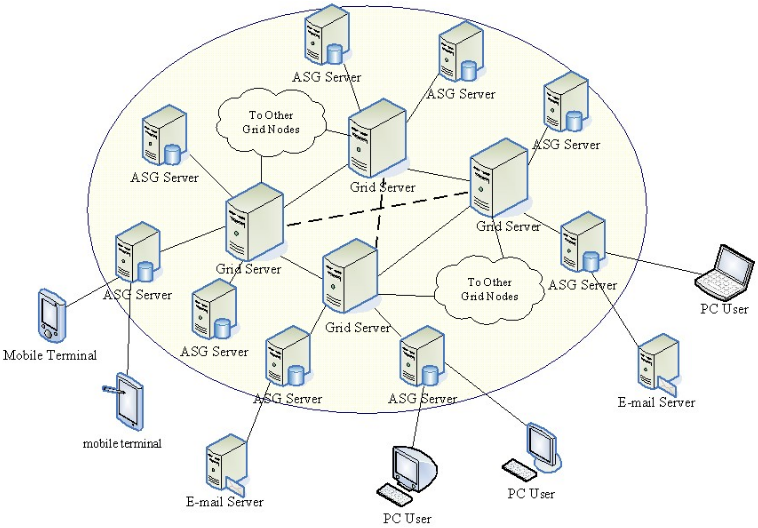

Fig. 1. The structure of Anti-Spam Grid

ASG Client includes the ultimate user of email service and the email servers that need service of spam filtering. When a new client takes part in the ASG, it suggests the request to a Grid Server. The Grid Server auditing the request and provides an authority certificate to the client and assigns an ASG Server who will serve the new client. ASG Client takes charge of the fingerprint abstracting. When a client send a fingerprint to ASG Server to query the amount of approximate email, the assigned ASG Server will responses the query and the new fingerprint will be stored in the global fingerprint repository. ASG Client also reports the local Bayesian knowledge to the assigned ASG Server.

The main work steps of Anti-Spam Grid are as followings:

- ASG Server publishes its service to someone Grid Server. Grid servers share their information each other.
- Once a ASG Client joined the ASG, it suggest request to a Grid Server.
- The Grid server assigns a ASG Server to serve it according to the rule of workload balance or serving nearby
- When the ASG Client suggesting connection request firstly to an ASG Server, the ASG Server should send the client's authority certificate to Grid Server to check it. If the certificate is qualified, it will be stored locally and the later checking will be implemented locally.
- ASG Client report email fingerprints and Bayesian knowledge to ASG Server.
- ASG Server returns the amount of approximate emails and other Bayesian knowledge to the client.

Since each ASG Server only stores the fingerprints of limited range, the query of approximate email may be processed in other ASG Server. The range of which fingerprint be stored in an ASG Server is assigned by a Grid Server. In order to increase the efficiency of approximate email query, every ASG Server maintains a fingerprint table. It only cost one hop to route the object ASG Server in approximate fingerprint searching.

The scalability of Anti-Spam Grid is perfect. Not only Grid Server but also ASG Server can join the system dynamically. When a new ASG Server join the system, it can burden apart of client and workload schemed by Grid Server.

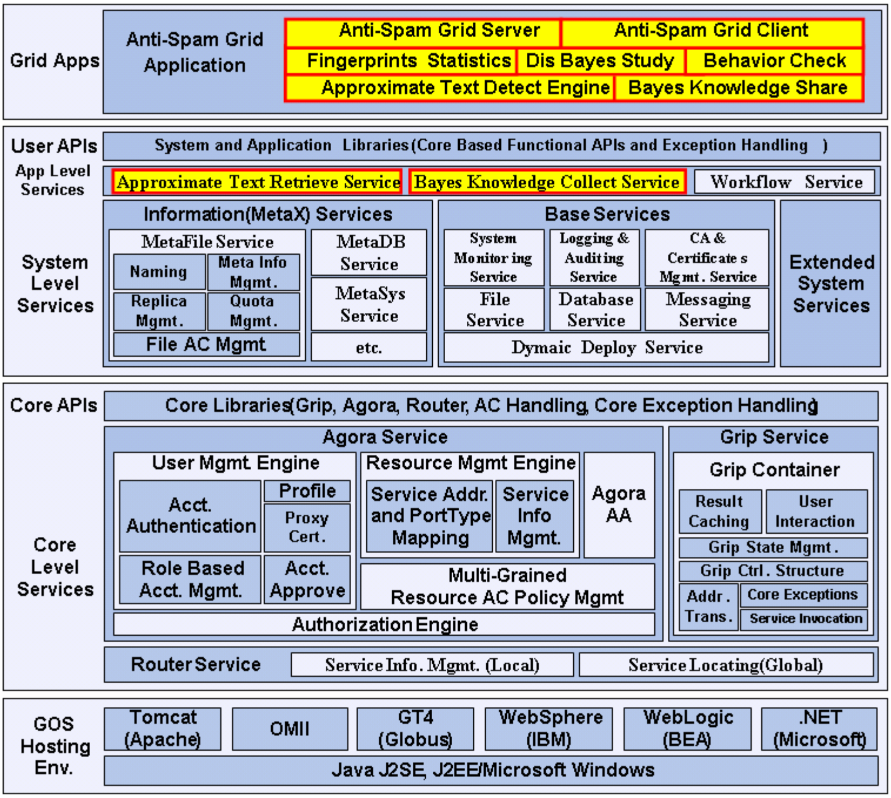

Fig. 2. The architecture of CNGrid-ASG

The proposed system is constructed on the CNGrid software version 2.0. The structure is shown in Fig. 2. We append the service of approximate text detection and the service of Bayesian knowledge collection and integrating, combined with the Information Service, System Monitoring Service, Logging Service, CA & Certificates Managements Service in the System Level and the User Management Engine in the Core Level, we realized the source management, user management and safe management.

Based on the CNGrid basis service and the extended service above, the system carried out Grid Server, ASG Server in Grid Application Lever. The system can filter spam in effect based on the integration of approximate email statistic technology, Bayesian knowledge sharing and real time communication protocol behavior analysis technology.

4 Simulation and the Result

There are many mature spam corpus for experiment of English spam filter such as Ling-Spam and PU corpus. In this paper we use a larger spam corpus, Genspam [13]. It includes 32332 spam and 9072 legit emails. In the spam corpus there are some spam are approximate. We used 90% of the corpus to train our Bayesian filter and use the other emails to simulate the process of email sending. In the simulate experiment, the amount of ASG users is from 10 to 100. We send a spam to *n* users, *n* is random from 1 to the user amount. We set the threshold as 10, when an email is receipted by users more than the threshold it will be judged as spam.

In classification tasks, two commonly used evaluation measures are accuracy (Acc) and error rate ($\text{Err} = 1 - \text{Acc}$):

$$Acc = \frac{n_{L \to L} + n_{S \to S}}{N_L + N_S}, \quad Err = \frac{n_{L \to S} + n_{S \to L}}{N_L + N_S} \tag{1}$$

N_L and N_S are the numbers of legit and spam emails to be classified. $n_{L \to L}$ is the number of spam emails that be classified as spam, and so on $n_{S \to S}$, $n_{L \to S}$, $n_{S \to L}$ can be deduced by analogy.

Accuracy and error rate assign equal weights to the two error types ($L \to S$ and $S \to L$). When selecting the threshold of the filter, but for users it is common believed that $L \to S$ is more costly than $S \to L$. To make accuracy and error rate sensitive to this cost, we adopt the cost-sensitive evaluation measures proposed by Androutsopoulos [14]: when a legitimate message is misclassified, this counts as λ errors; and when it is classified correctly, this counts as λ successes. This leads to weighted accuracy (WAcc) and weighted error rate (WErr =1−WAcc):

$$WAcc = \frac{\lambda \cdot n_{L \to L} + n_{S \to S}}{\lambda \cdot N_L + N_S}, \quad WErr = \frac{\lambda \cdot n_{L \to S} + n_{S \to L}}{\lambda \cdot N_L + N_S} \tag{2}$$

And the total cost ration (TCR):

$$TCR = \frac{WErr^b}{WErr} = \frac{N_S}{\lambda \cdot n_{L \to S} + N_{S \to L}} \tag{3}$$

In our experiment we set the λ as 9. The results are shown in Fig.3 and Fig. 4. For the Bayesian filter itself, the WAcc is about 91% and the TCR is about 2.5 and they are not change follow the user amount changing. For the ASG system, the WAcc and TCR are increasing with the in increase of users. In the experiment, the value of threshold is fixed. How to dynamic amend the threshold value is a problem should to research and experiment in our future work.

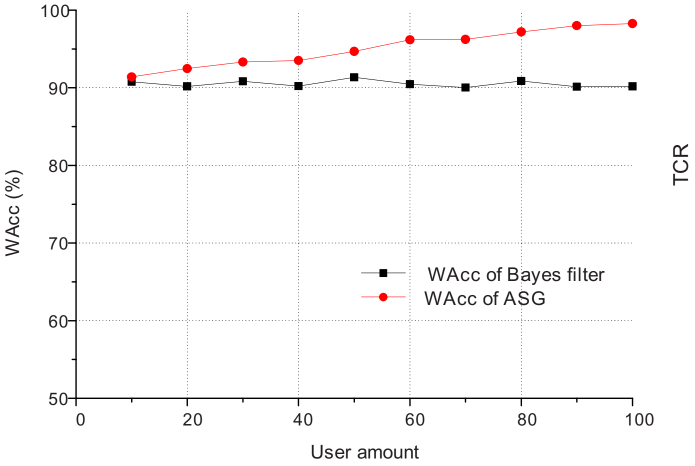

Fig. 3. WAcc of Bayesian filter and ASG

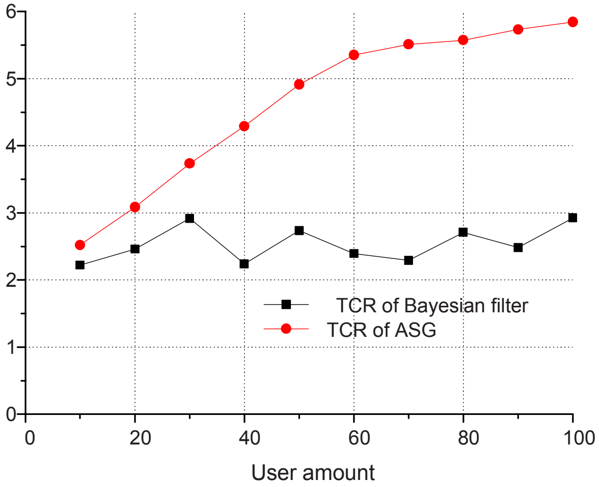

Fig. 4. TCR of Bayesian filter and ASG

5 Conclusions

We present design and evaluation issues of Anti-Spam Grid, an infrastructure dedicated to filter unsolicited bulk e-mails. Based on the CNGrid, the ASG users and servers can be dynamically added to the system. At the same time, the system can run properly whenever any ASG Server or Grid Server fails. So we can say the system reflects the core idea of the grid: virtual organizations. The result of experiment shows it is much more effective than contemporary anti-spam approaches.

References

1. Paul, N. C.and Monitor, C. S.: New strategies aimed at blocking spam e-mail. http://newsobserver.com/24hour-/ technology/story/655215p-4921708c.html.
2. Nelson, M.: Anti-spam for business and isps: Market size 2003-2008. Ferris research – analyzer information service report (2003)
3. Fallows, D.: Spam: How it is hurting e-mail and degrading life on the internet. Tech. Rep. 1100, PEW Internet & American Life Project (2003)
4. http://www.anti-spam.cn/ShowArticle.php?id=2713
5. Filters that Fight Back, http://www.paulgraham.com/ffb.html
6. Hird, S.: Technical Solutions for Controlling Spam. In proceedings of AUUG2002, Melbourne, September 2002.
7. ePrivacy Group, http://www.eprivacygroup.com/
8. Cohen, W. W.: Learning rules that classify e-mail.In: Proceeding of the AAAI Spring Symposium on Machine Learning in Information Access, 18-25, 1996
9. Spamassassin, http://au2.spamassassin.org/doc.html
10. Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E.: A Bayesian Approach to Filtering Junk E-Mail. AAAI Tech Rep. WS-98-05 (1998)
11. Ioannidis, J.: Fighting Spam by Encapsulating Policy in Email Addressed. 10th Network and Distributed System Security Symposium (2003)
12. Foster, I., and Kesselman, K. (ed.): The Grid 2: Blueprint for a New Computing Infrastructure, 2nd edition, Morgan Kaufmann, Nov. 2003, ISBN: 1558609334
13. http://www.cl.cam.ac.uk/~bwm23/
14. Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Paliouras, G., and Spyropoulos, C. D.: An Evaluation of Naïve Bayesian Anti-Spam Filtering. Proc. of Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona (2000)