Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 5 (2013) 220 – 227

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Parallel and Distributed Computing Systems

A Reputation-based Collaborative Approach for Spam Filtering

Wenxuan Shi[a], Maoqiang Xie[b]

[a]*College of Software, Nankai University, Tianjin 300071, China*
[b]*College of Software, Nankai University, Tianjin 300071, China*

Abstract

Spam and spam filters are contrarious components of a complex interdependent social ecosystem. Traditional spam filtering techniques or systems are usually designed and deployed individually that neglect the distributed and bulk characteristics of spam. This paper proposes a reputation-based collaborative anti-spam approach. This approach, adopting fingerprinting technique, evaluating reporters' trust, achieved better performance and robustness than the state-of-the-art in comparison experiments on several known email corpora.

Keywords: Spam filtering; reputation evaluation; collaborative; fingerprinting

1. Introduction

There is frequent and close mutual relationship between people and data on web, and such relationship offers some lawbreakers potential opportunity to send all kinds of spam information including spam email, spam poster, spam invitation, spam advertisement, etc. According to characteristic of spam information, there are many definitions about spam in practice, such as unsolicited and unwanted email, indiscriminate bulk email sent directly or indirectly, unsolicited commercial email, etc. In addition to wasting recipients' time to deal with spam, spam also eats up a lot of network bandwidth. Moreover, spam occupies vast resources of computation and storage and it is profoundly annoying clients and email service providers (ESP).

Consequently, many anti-spam techniques and systems, known as spam filters, have appeared associated with the appearance of spam. Such situation is called as spam ecosystem, that spam and spam filters are components of a complex interdependent system of social and technical structures. The anti-spam techniques

doi:10.1016/j.aasri.2013.10.082

can be categorized into two types: anti-spam techniques based on EK (expert knowledge) and anti-spam techniques based on ML (machine learning). Anti-spam techniques based on EK contain rule-based filtering, such as whitelists, blacklists, challenge-response, enhanced protocol, etc. There have been many spam filters based on EK, such as Sender Policy Framework (SPF) [1], Sender-ID [2], and Domain Keys Identified Mail (DKIM) [3], etc. Anti-spam techniques based on ML contain probability-based filtering, linear classifiers, Rocchio method, nearest neighbor method, logic-based method, data compression model, etc.

Most of recent spam systems are individual spam filters which are designed and deployed individually according to various clients and ESPs. Considering the distributed and bulk characteristics of spam, the better schema of anti-spam system should be implemented depend on multi-client or multi-ESP collaboration and has individuals sharing their judgments of legitimate email and spam. This paper proposes a reputation-based collaborative anti-spam approach, adopting fingerprinting technique, evaluating reporters' trust, and it outperformed the state-of-the-art in comparison experiments on several known email corpora.

2. Related work

In spam ecosystem, spam filters should pay close attention to two issues: one is how to protect recipients' privacy and information safety, and another is how to utilize the distributed and bulk characteristics of spam.

2.1. Fingerprinting

Fingerprint and fingerprint recognition are concepts originated from the area of biometric authenticate identity. After applying them in the area of information retrieval, fingerprint is defined as short tag for large object. Fingerprinting technique has the advantage that it can identify the same or similar duplicated documents with small partial variations by calculating and matching their hash values.

In spam ecosystem, spam information has produced massive waste to network resource and person's time, such as mass data transmission on internet and mass storage on servers. On the other hand, spam information has produced serious damage to web information safety and personal information privacy. Fingerprinting schema is derived from traditional research domain of data encryption and digital signature. This schema adopts certain encryption hash arithmetic by which to generate shorter content digests for large original messages in order to take the place of original information storage and transmission. Fingerprint functions may be seen as high-performance hash functions and there are two kinds of known algorithms: Rabin's algorithm and cryptographic hash functions [4].

2.2. Collaborative spam filtering

Collaborative spam filtering is more efficient strategy to content filtering where rather than employing someone or certain computers to attract and analyze spam, and rather than having different user train himself individual filters. The whole collaborative community works together with shared spam knowledge. Therefore a collaborative spam filter needs certain shared and efficient database where storing different user's judgment about which is spam and which not. At present there has been some collaborative spam filters on web, such as DCC, Vipul's Razor, Pyzor, Cloudmark, etc[5, 6, 7]. These methods have similar strategy of using shared knowledge. The DCC (Distributed Checksum Clearinghouse) system detects spam by computing spam checksums and querying the same checksums in a database of checksums. Vipul's Razor system filters out known spam by maintaining a catalogue server of signatures feedback by spam receivers. Pyzor and Cloudmark have different protocols of spam filtering similar to Vipul's Razor.

3. Reputation Evaluation

This paper proposes a collaborative approach for spam filtering based on reputation evaluation with weighing fingerprints and shared fingerprints database by means of which we can record, query, log, report and amend shared fingerprints, as shown in Figure 1.

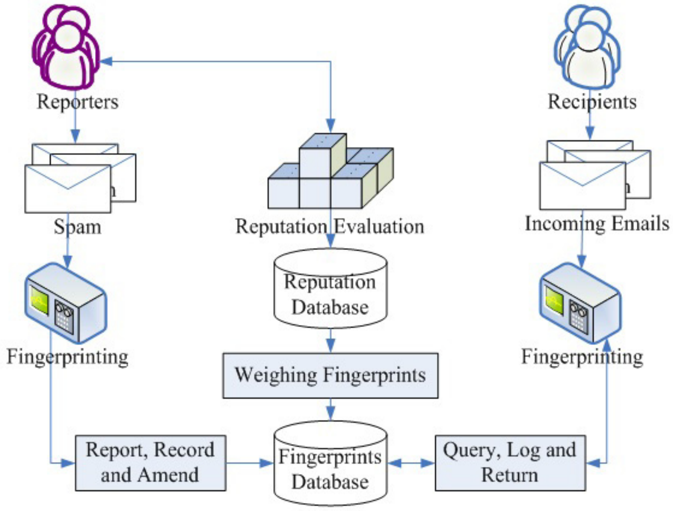

Fig. 1. Reputation evaluation with weighing fingerprints

3.1. Fingerprinting based on MIME-division

Multipurpose Internet Mail Extensions (MIME) is an Internet standard that extends the format of email to sup-port a variety of format content, such as text in multiple character sets, non-text attachments, and message bodies with multiple parts, etc. However, the traditional spam filtering techniques commonly just consider plain text content of email but ignore MIME Features, and only considered the text content, as well as some released email corpora to public which only contain text content extracted from original email body. In our work, the anti-spam system handles spam as follows:

1) Divide each incoming email into five subparts: email header, text/plain content of email body, text/html content of email body, embedded resources and attachments.

2) Generate weighing fingerprint for different MIME subpart.

3) Compute an indicator score for each fingerprint set to indicate how spammy the fingerprint set is.

4) Calculate a compound weighted score based on individual indicator score.

5) Make a decision of spam or non-spam to the incoming email by comparing the compound weighted score with certain predefined threshold value.

Define an incoming email as symbol m, after the processing of MIME-division, the email m is turned into a subparts set: $m = \{m^{[1]}, m^{[2]} \ldots m^{[5]}\}$ with the weighting vector $w = \{w^{[1]}w^{[2]} \ldots w^{[5]}\}$. Suppose the subpart $m^{[i]}$can be expressed as a feature vector $m^{[i]} = \{m_1^{[i]}, m_2^{[i]} \ldots m_n^{[i]}\}$. Suppose the generated fingerprint set of $m^{[i]}$ is $fig(m^{[i]}) = \{fig(m_1^{[i]}), fig(m_2^{[i]}) \ldots fig(m_n^{[i]})\}$. The compound weighted score $CW(m)$ could be calculated as follows:

$$CW(m) = \sum_{i=1}^{5} w^{i} \sum_{j=1}^{n} score(fig(m_j^{[i]})) \tag{1}$$

where $score(fig(m_j^{[i]}))$ is the indicator score of fingerprint $fig(m_j^{[i]})$ which can be trained, maintained and queried from certain fingerprints database. In the fingerprints database, we conserve every fingerprint information as a $< Fingerpr\mathrm{int}\_ID, Indicator\_Score, \mathrm{Validity}\_Datetime >$ triplet where $Fingerpr\mathrm{int}\_ID$ is a global unique hash value calculated by fingerprinting arithmetic, $\mathrm{Validity}\_Datetime$ is the life cycle or generated fingerprint.

3.2. Reputation evaluation

In our spam filtering system, we build a reputation evaluation method with weighing fingerprints and shared fingerprints database. By creating the shared fingerprints database, we can record new arrival fingerprints, query stored fingerprints, log users operations, report searching results and amend outdated fingerprints. To a reporter of the collaborative spam filtering system, we calculate the reporter's reputation value as follows:

$$\mathrm{Re}\, p(r,t) = \mathrm{Re}\, p(r,t-1) + \beta(r,t) \cdot j_{r,t}(m) \tag{2}$$

- $\mathrm{Re}\, p(r,t)$ is current reputation value of reporter calculated by last reputation and current feedback result multiply by weighing factor,
- $j_{r,t}(m)$ is reporter feedback result to certain email-m, which can be calculated as follows:

$$j_{r,t}(m) = \begin{cases} 1, & feedback \quad is \quad correct \\ -1, & feedback \quad is \quad incorrect \end{cases} \tag{3}$$

- $\beta(r,t)$ is weighing factor for different kinds of feedback rolls in order to balance reporter feedback result:

$$\beta(r,t) = \log(1 + roll(r) \cdot |\sum_{i=1}^{t} j_{r,i}(m)|)^{\sim} \tag{4}$$

where

- $roll(r)$ is pre-defined adjusting parameter for different feedback rolls defined by different source types of reporter.

3.3. Indicator score calculation

To judge which is spam and which not on the basis of feedback, the collaborative anti-spam approach based on reputation calculates an indicator score for each generated fingerprint after MIME-division and weighting-assignment as follows:

$$score(fig(m_j^{[i]})) = \sum_{u_k \in reporters(m)} (trust(u_k) \times odds(m_j^{[i]}))^{\sim} \tag{5}$$

where

- $score(fig(m_j^{[i]}))$, the indicator score of fingerprint $fig(m_j^{[i]})$, is the second element $Indicator\_Score$ value of the $< Fingerprint\_ID, Indicator\_Score, Validity\_Datetime >$ triplet which is stored in fingerprints database.
- $trust(u_k)$ is the reputation value of reporter u_k which can be configured and adjusted over time based on different reporter rolls defined as above($Rep(r,t)$).
- $odds(m_j^{[i]})$ is the probabilistic value of feature $m_j^{[i]}$ calculated based on IDF (inverse document frequency).

3.4. Spam detection

Spammers often control some puppet PCs and engaged servers to send mass spam. It is the purpose of spam filters to recognize these machines and senders. Spam filtering is the automatic processing to distinguish spam from non-spam between incoming emails according to specified criteria. After the steps of fingerprinting based on MIME-Division, and indicator score calculation based on reputation evaluation, we can generate a diverter for spam and non-spam as shown in Figure 2.

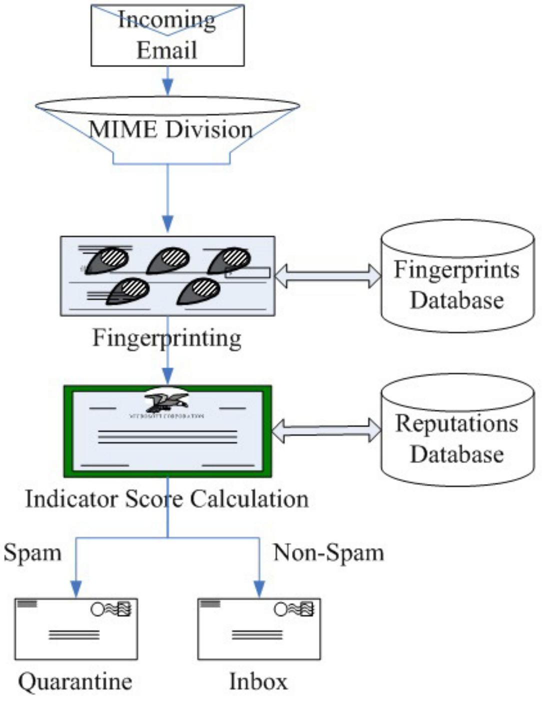

Fig. 2. Diverter for spam and non-spam

In order to predict the label of an incoming email and then put into appropriate email folders, we can execute the judgment as follows:

$$Label(m) = \begin{cases} Spam, & if \quad CW(m) > \tilde{t} \\ Ham, & if \quad CW(m) \le t \end{cases} \tag{6}$$

where t is a pre-defined judgment threshold of legitimate email and spam.

4. Experiments

In order to examine the performance of our approach (Collaborative Anti-Spam Scheme, CAS3 for short) proposed in this paper, we performed three groups of experiments based on three different corpora.

4.1. Experiment evaluation measure

We performed three groups of experiments based on three different corpora (Table 1) where the Handwork corpus is collected by our previous accumulation which contains whole email content, such as attachments, embedded resources, etc.

Table 1. Corpora used in comparison experiments

Corpus	Spam	Non-Spam	MIME
Ling-Spam [8]	481	2412	Subject, Text/Plain
Spam-Assassin [9]	1897	4150	Five Subparts
Handwork	4059	831	Five Subparts

In each group of experiments, we compare CAS3 with two known spam filters: SpamAssassin and Vipul's Razor. In comparison experiments, we draw certain Receiver Operating Characteristic curve to show different filtering effects between these methods.

4.2. Experiment evaluation results

The first group of experiments is simulated based on Ling-Spam corpus and the comparison results are shown in Figure 3.

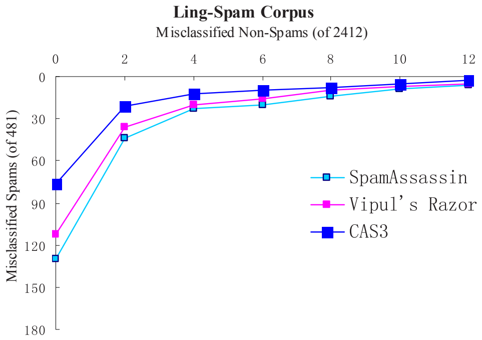

Fig. 3. ROC curve graph on Ling-Spam corpus

The second group of experiments is simulated based on Spam-Assassin corpus and the comparison results are shown in Figure 4.

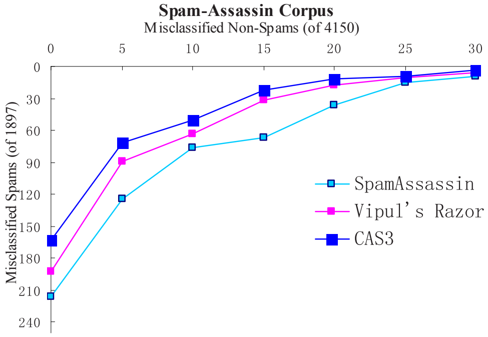

Fig. 4. ROC curve graph on Spam-Assassin corpus

The third group of experiments is simulated based on Handwork corpus and the comparison results are shown in Figure 5.

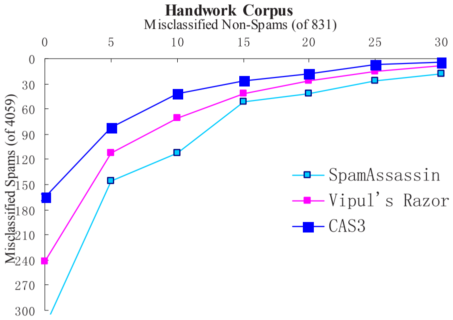

Fig. 5. ROC curve graph on Handwork corpus

5. Conclusions

In this paper, a reputation-based collaborative approach for spam filtering has been proposed that using the MIME features of email and adopts fingerprinting schema according to different subparts of email. Our approach achieved better performance and robustness than current popular filtering methods on several email corpora.

Corresponding Author:

Wenxuan Shi, shiwx@nankai.edu.cn, 086-13920561100

References

[1] M. Wong and W. Schlitt. 2006. Sender Policy Framework (SPF) for Authorizing Use of Domains in E-

mail, Vol. RFC 4408.
[2] J. Lyon and M. Wong. 2006. Sender-ID: Authenticating E-mail RFC 4406, Internet Engineering Task Force.
[3] B. Lieba and J. Fenton. 2007. DomainKeys identified email (DKIM): Using digital signatures for domain verification. in CEAS 2007: The Third Conference on Email and Anti-Spam.
[4] wikipedia.org. 2012. Fingerprint (computing). http://en.wikipedia.org/wiki/Fingerprint_(computing). Andrew G. West, Avantika Agrawal, etc. 2011. Autonomous link spam detection in purely collaborative environments. In WikiSym 2011: The 7th International Symposium on Wikis and Open Collaboration. Network Security, pp. 15–17.
[5] Wenxuan Shi, Maoqiang Xie, etc. 2011. Collaborative Spam Filtering Technique Based on MIME Fingerprints. In WCICA 2011: The 9th World Congress on Intelligent Control and Automation. Washington, DC, USA: IEEE Computer Society. 2011.
[6] Wenxuan Shi, Maoqiang Xie, etc. 2011. Cooperative Anti-Spam System Based on Multilayer Agents. In WWW 2011: The 20th International World Wide Web Conference. NY, USA: ACM. 2011. 415~420.
[7] Stason.org. 2006. Anti-SPAM Techniques: Collaborative Content Filtering. http://stason.org/articles/technology/email/junk-mail/collaborative_content_filtering.html.
[8] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, George Paliouras, and C.D. Spyropoulos. 2000. An Evaluation of Naive Bayesian Anti-Spam Filtering. in Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona, Spain, pp. 9-17.
[9] Spamassassin.org. 2003. The Spamassassin Public Mail Corpus. http://spamassassin.apache.org/publiccorpus.