available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Using header session messages to anti-spamming

*Chih-Chien Wang**, *Sheng-Yi Chen*

Graduate Institute of Information Management, National Taipei University, Taipei 104, Taiwan

ARTICLE INFO

Article history:
Received 7 December 2005
Revised 11 September 2006
Accepted 22 December 2006

Keywords:
Spam
Unsolicited email
Junk mail
Email address
Filter

ABSTRACT

The Internet is popular, with email use functioning as the major Internet activity. However, spam has recently become a major problem impeding the use of email. Many spam filtering techniques have been implemented so far. Most current anti-spamming techniques filter out junk emails based on email subjects and body messages. Nevertheless, subjects and email contents are not the only cues for judging spam. This investigation presents a statistical analysis of the header session messages of junk and normal emails, and explores the possibility of utilizing these messages to perform spam filtering. The message head session, including the sender's mail address, receiver's mail address and time, which is of little interest to most users, also provides further information for anti-spamming purpose. A statistical analysis is undertaken on the content of 10,024 junk emails collected from a Spam Archive database, and 599 regular emails in company with 635 solicited listserv or commercial emails contributed by volunteers. Content analysis results demonstrate that up to 92.5% of junk emails are filtered out when utilizing the message-ID, mail user agent, and sender and receiver addresses in the header session as cues. Additionally, the proposed approach may induce a low block error rate for normal emails for the sample utilized in this investigation. This low rate of over-block errors is a significant merit of the proposed anti-spamming approach. The proposed approach of utilizing header session messages to filter out junk emails may coexist with other anti-spamming approaches. Therefore, no conflict arises between the proposed approach and existing spam prevention approaches.

© 2007 Elsevier Ltd. All rights reserved.

1. Introduction

The Internet has been growing vigorously in recent years, having evolved into a necessity of daily life for many individuals. Using email is thus the major activity when surfing the Internet. However, the problem of spam flooding user mailboxes wastes users' time in reading them, and costs billions of dollars in wasted bandwidth and disk storage space (Cournane and Hunt, 2004). The existence of junk emails is clearly an irritation when using the Internet. This problem will increase if without further research into anti-spamming techniques.

Postel (1975), an Internet pioneer, recognized the potential problem of junk email as long ago as November 1975, and proposed Requests for Comments (RFC) 706 describing the problem of junk email. The first junk email was sent on the Arpanet in 1978 (Hinde, 2003). ACM former president Peter J. Denning published the first article about junk email in 1982, in Communications of ACM, predicting that junk email would abound (Denning, 1982). The term "spam" became widespread in April 1994, when two American lawyers, Canter and Siegel, hired a programmer to write a simple script to post their advertisement to every newsgroup board on USENET in order to propagate their U.S. "green card" lottery

* *Corresponding author.*
E-mail address: wangson@mail.ntpu.edu.tw (C.-C. Wang).
0167-4048/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cose.2006.12.012

service (Cranor and LaMacchia, 1998). This event seems to be the beginning of the use of the term ''spam''.

Many studies aimed at anti-spamming, including Damiani et al. (2004) presented peer-to-peer architectures between mail servers to collaboratively share knowledge about spam filters; Jung and Sit (2004) focused on DNS blacklists; Sahami et al. (1998) employed the Bayesian approach to filter junk emails; Rigoutsos and Huynh (2004) developed a Chung-Kwei algorithm based on the genetic algorithm to implement a system to recognize junk email; Golbeck and Hendler (2004) presented an email scoring mechanism based on a social network augmented with reputation ratings; Leiba and Borenstein (2004) found that no particular technique fully solves the spam problem, but that different techniques excel in different ways. Hence, they proposed the application of multiple techniques in several layers in order to improve filtering efficiency.

Conventional techniques mentioned above are unable to filter out all spam emails. Spammers might find a way to avoid being filtered out. Junk emails may contain faked sender's email address, sent in many batches of with a few messages each, and with subjects irrelevant to the mail content, to avoid being filtered by anti-spamming techniques. Therefore, more effort is required to enhance current anti-spam techniques.

Additionally, some filtering techniques may filter out many spam messages, but also block many normal emails. For instance, some anti-spamming mechanisms might filter out emails containing the word ''adult'', even though they might be solicited emails for ''continuing adult education'' sent to or by scholars majoring in a particular field. Furthermore, a manager might want to email a message to all employees, but such bulk emails might also be filtered by the anti-spam mechanism. Anti-spam mechanisms have high over-blocking rates in both these scenarios.

Over-blocking of normal emails as spam may cause loss of emails and make the email service unreliable. The over-blocking problem can cause trouble for users in their daily life or work. Internet users who encounter over-blocking may not trust email any more. Under these conditions, people sending out emails have to confirm the delivery of email to the receiver. This over-blocking problem makes email lower the convenience of email.

Most current anti-spamming techniques are memory-based approaches that filter junk emails out based on email subjects and body messages. However, subjects and email contents are not the only cues for evaluating spam. Each email has a header session composed of several headers about factors such as the sender's mail address, receiver's mail address, mail servers, client email software, message identity number and time stamp. These messages found in the header session may be used as cues for spam filtering.

Due to the imperfection of existing anti-spamming techniques, additional effort is necessary to find novel anti-spamming approaches. This investigation presents a statistical analysis of header session messages of junk and normal emails to explore the possibility of utilizing these messages to filter spam. In addition to the content and subject, email has a header session containing some fields about the messages. Past anti-spamming techniques typically utilized email subjects or contents as cues for anti-spam filtering, and disregarded the information in the header session.

The header session stores information about email delivery. Internet users generally are not concerned with the information in the header session, except for the sender email address for replying to email, and the time stamps for sorting emails. Most users are not interested in other information. However, these headers that do not interest users may be effective cues for anti-spam filtering.

A statistical analysis was conducted of message-ID, mail user agent, sender and receiver addresses in the header session of junk and normal emails. The email addresses of the sender and receiver identify the sender and receiver. Spammers typically send junk emails with invalid sender addresses to avoid possible accusation and suspension of email service by their Internet Service Providers. Additionally, most junk emails are sent out in bulk. Spammers typically put receivers' emails in the ''BCC'' (Blind Carbon Copy) field, and do not reveal the names of the real receivers to prevent receivers from knowing that this email is sent out to a large number of recipients. Restated, the receivers' email addresses would not been found in the ''To'' and ''CC'' (carbon copy) receiver fields in most junk email.

The X-Mailer field of the header session of an email message indicates the client software or mail user agent (MUA) used to send it. Normal personal communication emails are sent out via client software such as Outlook, Outlook Express, and Lotus Notes. However, junk emails are sent out using bulk email software programs, which either do not disclose their software name in the X-Mailer field or place random meaningless characters there.

The message-ID field is generated by either the MUA or the first mail transfer agent (MTA) through which the message passes, and thus uniquely identifies an email message. Most spammers would fake the part value of message-ID field, so that the domain name of the sender address does not match the domain part of the message-ID. Spammers hope to avoid revealing the real domain name of the MUA or the first MTA through which the message passes, and therefore add a fake domain name to message-ID field.

The remainder of this paper is structured as follows. The following section review anti-spamming mechanisms and the efficiency of anti-spam techniques. The next section describes the research design, and details the results of content analysis. Finally, this paper discusses the possibility of utilizing header session messages as cues for anti-spam filtering. The analytical results of this study demonstrate that these header session messages might be valuable for screening out junk emails.

2. Anti-spamming approach

The accelerated increase in spam traffic is a significant problem to end users and businesses. Several techniques have been proposed to filter spam out. However, these anti-spamming methods have so far had a very limited effect on the amount of spam received. Various spam filtering techniques are listed below.

2.1. Bulk email filtering

Bulk email filtering is the easiest means of filtering out bulk email for a mail server. This approach is based on the assumption that junk emails are typically sent to a large number of recipients, so the system administrator only needs to set an upper limit for the number of recipients of every email at the mail server. However, a spammer can easily avoid this filtering technique by disconnecting and reconnecting after sending a particular number of emails. Some bulk mail programs can avoid this filter simply by constantly changing the email subject information. Additionally, this method may mistakenly filter out important email going to a large number of recipients. The key determinant of this method's efficiency is the recipients' upper limit setting. If the setting is too strict, then this filtering technique can filter out many normal emails. By contrast, a setting that is too lax would not filter out junk email.

2.2. Filtering by keyword

Keyword filtering is the most frequently utilized anti-spamming technique. There are two approaches for applying this method. The first approach which is easy but inefficient, filters out all emails in which a particular keyword appears in the subject or content. ''Sex'', ''On Sale'' and ''Get Rich'' are frequently used keywords. However, this method is inefficient, since while the spammer can avoid using these keywords, enabling the spam to be sent, while normal emails containing these keywords are by mistakenly treated as junk emails, and filtered out.

The second approach of filtering out junk emails is based on results of machine learning. This approach is more complex and more efficient than the above approach. This approach determines the keywords using a machine-learning algorithm, and calculates the frequency of all keywords to discriminate between junk and normal emails. Anti-spam methods that are based on this approach include the RIPPER rule learning algorithm (Cohen, 1996; Provost, 1999), naïve Bayesian classifier (Sahami et al., 1998; Schneider, 2003; Sinclair, 2004), support vector machine (Drucker et al., 1999; Kolcz and Alspector, 2001; Gordon and Hongyuan, 2004; Woitaszek et al., 2003), centroid based (Soonthornphisaj et al., 2002), decision trees (Carreras and Marquez, 2001) and memory-based filtering (Androutsopoulos et al., 2000). This method is based on the assumption that the subject or body sector of junk emails may contain specific words. However, it does not work perfectly, since it might filter out normal emails simply because they include too many words on the keyword list for junk emails.

2.3. Blacklist

This approach exploits a blacklist database to block specific email addresses or domain names. Such databases are available on the Internet, such as DNSBLs (http://dsbl.org/main) or SORBs (http://www.us.sorbs.net). If such a database is updated frequently, then it can reliably filter out certain known spammers' addresses. However, a serious problem with a blacklist is that it would filter out normal emails if it contains any normal email address or domain name. Furthermore, spammers typically leave random assigned faked sender addresses which are not on the blacklist. The blacklist approach cannot function well if sender addresses are faked.

2.4. Whitelist

A whitelist is designed to avoid filtering normal emails out. Whitelists gather permitted email addresses or domain names, and often collaborate with blacklists. Blacklists block illegal email addresses, while whitelists allow normal emails to pass through. A whitelist is difficult to maintain perfectly than that a blacklist. Including all normal email senders on a whitelist is an impossible task.

2.5. Sender address validity

Crocker in 1982 advised the Requests for Comments (RFC) 822 that there is a standard for the format of ARPA Internet text messages, which describes the format of email. According to RFC 822, each email message must include the ''From'' field, containing the address of the sender who wishes this message to be sent. Each email should include at least one sender, and must include the ''From'' field. Although RFC 822 mandates the existence of the ''From'' header in emails, the sender address can be invalid or faked by spammer. The spammers fake the senders' address to avoid being accused of sending junk email and breaking the law. Additionally, most Internet Service Providers (ISPs) or email service providers suspend the use of spammers' email address, or refuse to relay the emails sent by spammers. Wang (2004) concluded that 60.3% of junk emails provide invalid sender addresses. Consequently, the validity of sender address left in the ''From'' header session may be a cue for spam filters.

2.6. Receiver address as cue

RFC 822 also specifies that the receiver addresses be listed in the ''To'', ''CC'', or ''BCC'' fields. Each email must have at least one receiver. The ''To'', ''CC'', or ''BCC'' headers are applied to inform the recipients of this email where it has been sent. The ''To'' field contains the identity of the primary recipients of the message and ''CC'' standing for carbon copy, contains the identity of the secondary recipients of the message. Thus, the function of the ''To'' and ''CC'' fields is very similar. However, ''BCC'' differs from ''To'' and ''CC''. ''BCC'', which stands for Blind Carbon Copy, contains the identity of additional recipients of the message. The contents of this field are not included in copies of the message sent to the ''To'' and ''CC'' recipients. Wang (2004) found that only 7.2% of spam has the recipient's address in the ''To'' or ''CC'' fields, and that spammers typically place the recipient's address in the ''BCC'' field for the receiver address to avoid revealing that it has been sent in bulk. Therefore, the presence of the recipient's address in the ''To'' or ''CC'' header session may be a cue for spam filters. However, this approach may filter out normal email in which the recipients are listed in the ''BCC'' field. Email that lists senders' address in ''BCC'' field is not necessarily spam, although it has a high possibility of being spam.

2.7. *Mail user agent as cue*

In the header session, RFC 822 stipulates the use of X- at the beginning of a field name to indicate that the field is an extension. The X-Mailer field indicates what email client program or mail user agent (MUA) was used to generate the email. Although this field is not required in the header session, most MUA developers generally program their software to add an appropriate X-Mailer field to all out-bound emails. Most junk email programs either do not include the X-Mailer field, or place a random value in it, according to observations in this study. Therefore, the X-Mailer field may be a cue for spam filtering. The most frequently used MUAs, such as Outlook Express, MS Outlook, Lotus Notes and Eudora, correctly mark the X-Mailer field of all out-bound emails. By contrast, an in-bound email where the X-Mailer field value is null or contains a meaningless randomly assigned value is probably junk email. A normal MUA for sending bulk email would not fake the X-Mailer field, because feigning it as Outlook Express, MS Outlook, Lotus Notes, or some other MUA software may violate the trademark law. Bulk email software developers do not violate any law in most of cases. However, if they feign their program as other MUA software, they violate the trademark law. Consequently, the X-Mailer field can be a cue for spam filtering.

2.8. *Message-ID as cue*

The unique message identifier in the header session is generated by the MUA, or by the first MTA through which the message passes if the MUA did not assign one. This identifier is intended to be machine-readable, and is not necessarily meaningful to humans. The value of this message-ID field comprises two parts separated by an @ symbol. The left side contains a string of characters to uniquely identify the message on the machine where it was generated, and while its format depends on the email software, it is typically based on the date and time. The right side specifies that machine or domain name (Johnson, 1999: pp. 32–33). Most spammers would fake this domain value, such that the domain of sender address does not match the domain part of the message-ID, to avoid possible suspension of their Internet access. Spammers hope to avoid revealing the real domain name of the MUA or the first MTA, and therefore add a false domain name. Consequently, this condition provides a cue for deciding the possibility if an incoming email is a junk mail.

3. Efficiency of anti-spam techniques

Two classes of error need to be addressed when deriving the effectiveness of spam filtering techniques.

- Under-blocking occurs when an email that should be filtered out is not blocked.
- Over-blocking occurs when solicited normal email that should not be filtered out is blocked.

Bad anti-spamming techniques do not block some junk emails, and block some normal emails. These two error rates can be determined using formulae presented by Resnick et al. (2004), which are listed below.

$$\text{Under-blocking errors} = \text{unblocked junk emails}/(\text{blocked junk emails} + \text{unblocked junk emails})$$

$$\text{Over-blocking errors} = \text{blocked normal emails}/(\text{unblocked normal emails} + \text{blocked normal emails})$$

A good filtering technique minimizes both these error rates. However, under-blocking and over-blocking errors have different levels of importance. Over-blocking is a more significant problem than under-blocking for most email users. Unblocked junk emails only cause users to have to spend time in deleting them. However, over-blocked normal emails normally cannot be recovered. Thus, users would lose some important messages, possibly resulting in communication problems for work or daily life. If email users are aware of the possible risk of over-blocking, they have to ask the recipients to confirm receipt of their email messages. This may be inconvenient to email users, lower the reliability of emails.

Under-blocking and over-blocking are benchmarks for the effectiveness of anti-spam filtering techniques. Most previous studies have concentrated on under-blocking errors, although spam filters need to lower both error rates. For instance, Ahmed and Mithun (2004) reported that they could filter out 80% of junk email, but do not mention the over-blocking error rate. Shih et al. (2004) reported under-blocking errors as low as 7%, but did not mention the over-blocking error. Additionally, studies by Androutsopoulos et al. (2000) and Soonthornphisaj et al. (2002) did not mention the over-blocking error.

4. Content analysis for junk and normal emails

Content analysis was performed involving 10,024 junk emails, collected by Spam Archive (http://spamarchive.org) during a two-month period, and 599 normal emails and 635 solicited listserv or commercial emails, contributed by three volunteers over a one-week period. A high proportion of junk emails was utilized in this study to reflect the fact that most emails received by users are junk at present or in the near future. Spam Archive has collected a large number junk emails donated by end users, and is a well-known large spam repository for developing anti-spam tools. Normal emails are defined as emails for personal communication. Junk emails are unsolicited emails, and are typically sent in bulk and for commercial purposes. However, some emails sent in bulk are solicited by users. Listserv email is a typical case of this category. People may join a listserv email list, discussion board, or online community in order to receive emails. Additionally, users may subscribe to retail websites to receive updated sale messages. Such listserv and commercial emails should be treated as solicited, even though most of them are sent in bulk.

Content analysis was utilized to examine the possibility of using header session messages as cues to discriminate between normal and junk emails. The sender address, recipient address, message-ID and MUA fields in the header sessions of normal and junk emails were utilized for this purpose.

4.1. *Sender addresses*

The validity of the sender address was checked for all normal and junk emails collected by this study. Sender addresses were checked via a Domain Name Server (DNS), which checked for existence of the mail server, and Simple Mail Transfer Protocol (SMTP) was adopted to verify the existence of the mail account. A sender's email account was checked by sending an email if the SMPT servers refused to respond to email address validity checks. Figs. 1 and 2 illustrate the results of sender addresses checking. Of the 10,024 junk emails collected, 6664 (66.48%) had invalid (or non-existing) sender addresses. Of the collected 635 solicited listserv or commercial emails, only 28 (4.41%) had invalid sender addresses. Moreover, none (0%) of the normal email collected had an invalid address.

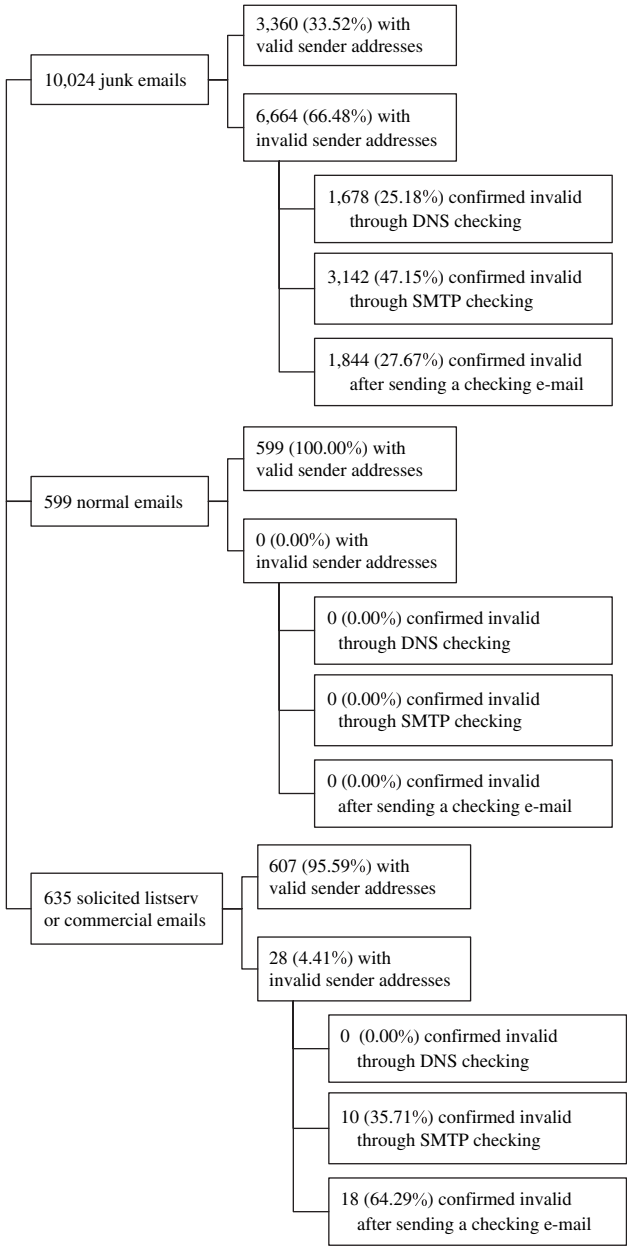

Fig. 1 – Sender address checking.

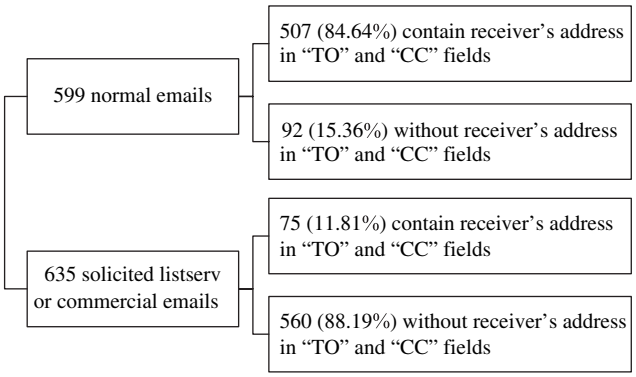

Fig. 2 – Receiver addresses.

The results of sender address validity checking indicated that sender addresses may be a cue for lowering the over-block rate. Fig. 1 demonstrates that all normal emails had valid sender addresses. Therefore, filtering out all emails with invalid sender addresses does not cause normal email to be mistakenly blocked out. Simply filtering out email without valid sender addresses may block out 66.48% of junk emails. However, this valid sender address rule would filter out 4.41% of solicited listserv or commercial emails. These filtered emails must be treated as over-blocked emails if solicited listserv and commercial ones are considered as normal. Nevertheless, the over-block rate is zero if users think that filtering out listserv and commercial emails is acceptable. Users can lower this over-block rate simply by including the senders' address of solicited listserv and commercial ones in the whitelist.

4.2. *Receiver addresses*

The presence of recipient address in the "To" or "CC" header session may be a cue for spam filtering, since spammers typically place the recipient address in the "BCC" field to avoid revealing that junk email is sent in bulk (Wang, 2004).

The junk emails collected by spamarchive.org do not include their recipient addresses. Therefore, this study could not analyze the recipient addresses of junk email. Fig. 2 illustrates the content analysis results of recipient addresses of normal and solicited listserv and commercial emails.

The content analysis shows that 84.64% normal emails contain receivers' address in "To" and "CC" fields, while 15.36% of normal emails do not have the recipients' addresses in "To" or "CC" fields. Additionally, 44.26% of junk emails and 11.81% of solicited listserv and commercial emails did not have the recipients' addresses in the "To" and "CC" fields, along with 55.74% of junk emails and 88.19% solicited listserv and commercial emails.

The recipient address could be utilized as a cue for judging normal emails. If an email has the recipient's address in the "To" or "CC" field, then it has a high probability of this email being a normal email. However, this should only be a supplementary cue, since some normal emails deliberately have the receivers' email addresses in the "BCC" header session.

Solicited listserv and commercial emails have high block-out percentages because they are mostly sent in bulk. The

recipient addresses of bulk emails can be expected to be placed in the "BCC" rather than "To" or "CC" fields. Over-blocking some normal emails (15.36%) and most solicited listserv and commercial emails (88.19%) is a side effect of using the recipient addresses to filter out emails. However, because most normal emails have the receiver addresses in the "To" or "CC" fields, the receiver addresses may still be additional cues for identifying normal emails. Additionally, the recipient addresses may still be valuable for spam filtering if over-blocking of solicited listserv and commercial emails is acceptable for users.

4.3. *Mail user agent*

Mail user agents (MUAs) are email client programs that used to write emails. Most normal emails have X-Mailer messages to indicate the MUA that sent them, although this is not required. However, most junk emails do not include the X-Mailer field, or include it with randomly assigned values.

Fig. 3 illustrates the content analysis results of the X-Mailer field, revealing that only 1.73% of junk emails were sent by bulk email programs. Additionally, no X-Mailer messages were found in 58.21% of the junk emails. Junk emails (5.69%) had randomly assigned values for the X-Mailer field while 2.03% junk emails were sent by rarely used MUAs.

Furthermore, most normal emails (52.96%) were sent using popular MUA program, as revealed by Fig. 3. Only 2.44% normal emails were sent by bulk email programs, while 0.87% were sent by infrequently used MUA, none with randomly assigning X-Mailer value, and 43.73% lacked an X-Mailer message.

The content analysis results indicate that MUA could be employed as a cue for normal email judgment. An email is likely to be normal if it is sent by a frequently used MUA. By contrast, the possibility that an email is junk is high if it is sent using a bulk email program, or the X-Mailer messages are randomly assigned. However, this should be a supplementary cue only, since normal emails transmitted by web mail interfaces may leave the X-Mailer field blank.

4.4. *Message-ID*

The message-ID is generated by the MUA, or by the first MTA through which the message passes if the MUA did not assign one for it. This may be utilized as a judgment cue for normal email. Most spammers hope to avoid revealing the real domain name of the MUA or the first MTA, and therefore add non-existing domain names. Therefore, if an email's sender address matches the domain name specified in the message-ID, then the probability that it is a normal email is high. However, an email whose message-ID does not match the sender address is not necessarily junk, since the message-ID may be assigned by the MUA instead of an MTA. If message-ID is assigned by MUA such as Outlook Express and some other MUA, then the computer name rather than the email domain name is placed on the right side of message-ID. Additionally, the authors' observation indicates that some web mail systems place software or computer names rather than domain names on the right side of message-ID, also makes preventing matches between the message-ID and sender address.

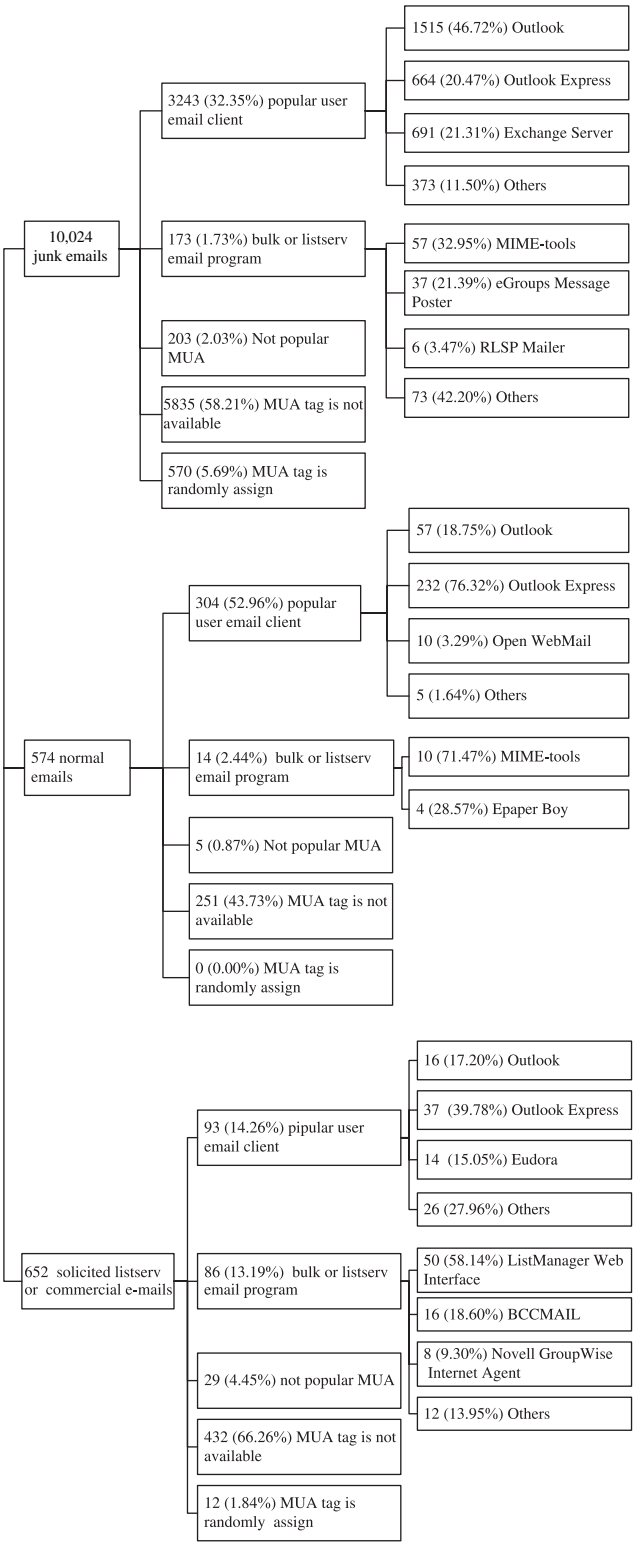

Fig. 3 – Mail user agent, MUA.

Fig. 4 presents the analytical results of matching between the message-ID and sender address. The results show that the message-IDs did not match the sender addresses for 82.66% of junk emails. Meanwhile, only 11.02% message-IDs of normal emails in study did not match the sender addresses.

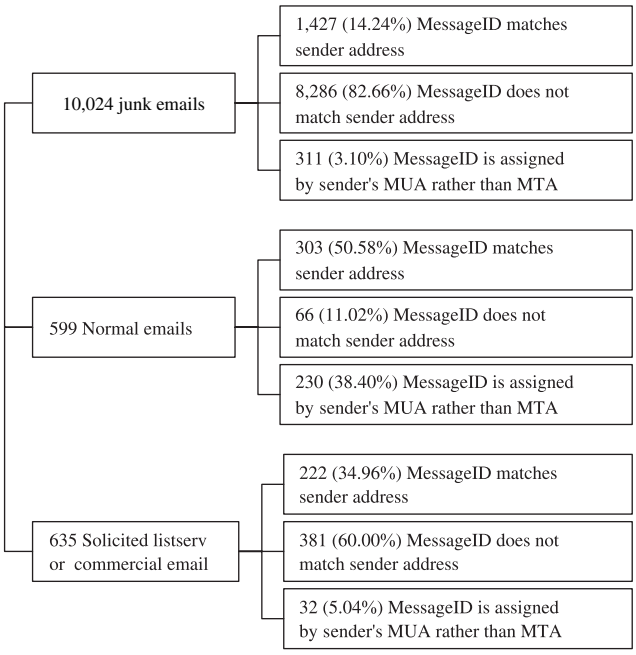

Fig. 4 – Match between Message-ID and sender address.

5. Using header session message to block spam

As mentioned above, sender address, receiver address, MUA and message-ID can be used as cues for anti-spamming techniques. If the sender address is invalid, then the email cannot be normal according to content analysis results, as revealed in Fig. 1. If an email does not contain the receiver address in either the "To" or "CC" fields, then it is highly likely to be junk, as demonstrated in Fig. 2. In addition, Fig. 3 indicates that the possibility of an email being sent in bulk is high if the MUA is a bulk email program, or MUA tag is not available or randomly assigned. In this situation, the email may be junk and should be filtered out. Furthermore, if the message-ID does not match the sender address, then the possibility is high that the message-ID is assigned by an MUA rather than an MTA, or that the message-ID is faked to avoid possible tracing. Some frequently used MUA programs, such as Outlook Express and some web mail programs, assign message-IDs to all sent out emails based on their own rules, and not relative to sender's address – so that the message-ID would not match senders' address. However, if an email is not sent by these MUAs, then the message-ID should match sender's address. If no match is found, then the probability of the email being junk is high, as demonstrated in Fig. 4.

The sender address, recipient address, MUA and message-ID cannot be used for spam filtering without considering other factors. Simply filtering out an email without the recipient address in the "To" and "CC" fields may cause over-blocking of some normal emails that are sent intentionally by Blind Carbon Copy. Similar situations are also found in using MUA and message-ID as anti-spamming cues. A user wanting to send a personal message to friends in a large batch may utilize a bulk email MUA. Filtering out emails sent by a bulk email MUA will over-block such emails, which should be treated as normal rather than spam. The same situation occurs for the message-ID. An email may be normal even if its message-ID does not match the sender's domain name. Some SMTP components, modules and libraries, for website programming do not arrange the message-ID tag thoroughly. The message-IDs in the messages sent using these MUAs do not match senders' addresses, even if they are normal emails.

This study develops a combined anti-spamming judgment approach since a single item consisting of items such as sender address, receiver address, MUA, and message-ID cannot be used as the sole cue for anti-spamming. This anti-spamming approach judges emails as normal or spam according to four cues, i.e. sender address, receiver address, MUA, and message-ID. An email will be judged as normal if it conforms to the rules of normal email. Again, it will be judged as junk when it conforms to the rules of junk. If an email conforms to neither junk nor normal, the email will be classified as indeterminate. To filter or not to filter this kind email should be the user's personal choice.

Table 1 illustrates the anti-spamming approach that this study proposes. The possibility is high of an email not being spam if it comes with a valid sender address, the MUA is not a frequently used bulk or automated email program, and the message-ID matches sender address or is assigned by the MUA rather than the sender's MTA.

However, if an email is probably sent by a bulk email program, message-ID does not match sender address, is a carbon copy to receivers, and has an invalid sender addresses, then it has a high probability of being junk.

The qualifications for all four spam rules mentioned in Table 1 are too strict. An email could still be spam if it does not comply with all four rules mentioned in Table 1. This study presents a formula that considers an email as spam if it either has an invalid sender address, or matches two of the remaining three rules (rules 2, 3 and 4) in Table 1. Rule 1, invalid sender address, is considered to be sufficient to filter out an email, although it would mistakenly filter out some listserv emails. Any listserv emails cannot reasonably be sent with an invalid sender address even they are sent upon subscribers' request.

Table 2 presents the spam filter results. The junk email filtering rate was 79.11% when the emails judged as spam were filtered out. However, this rate would rise by 13.39 percentage points to 92.50% if emails that are judged as indeterminate are filtered out, and only emails that are judged as normal are kept. However, the proposed anti-spamming approach filtered out 88.50% of solicited listserv and commercial emails, rising by 2.76 percentage points to 91.26% when it filtered out emails judged as indeterminate. Most solicited listserv and commercial emails look like junk emails. The proposed anti-spamming approach would filter out most of them, 88.50%, as revealed in Table 2.

Table 3 summarizes the anti-spam efficiency of the proposed approach. A user who hopes to filter out as many junk emails as possible can choose to keep only emails judged as normal, and filter out both spam and indeterminate emails. This process can be called 'stick' filtering. However, some normal email is judged as indeterminate as indicated by Table 2. Users have to accept the risk of mistakenly filtering out

Table 1 – Anti-spamming approach using sender address, receiver address, MUA, and message-ID

Judgment	Approach	Rules
Judged as normal emails	Do not filter out emails with all three characteristics.	Normal email has the following characteristics. 1. Valid sender address. 2. MUA is not frequently used bulk or automatic email program. 3. Message-ID matches sender address, or is assigned by MUA rather than sender's MTA.
Judged as spam	Filter out emails that match spam rule 1; or match any two of rules 2, 3, and 4.	Spam has the following characteristics. 1. Invalid sender address. 2. Email is not 'To' or 'CC' to recipient. 3. Sender's MUA is a bulk email program, or the MUA tag is not available or is randomly assigned. 4. Message-ID does not match sender address.
Judged as indeterminate	Neither normal or spam emails.	Neither normal nor spam emails.

normal emails. Conversely, a user may select a safe strategy and filter out only emails that are judged as spam. In this situation, all normal and indeterminate emails are kept. This can be called a 'slack' filter, since it only blocks confirmed spam.

The over-block error rate is the rate at which normal email, which should not be filtered out, is blocked. The over-block errors rate is zero if a slack filter is adopted. Therefore, no side effect occurs if a slack filter is adopted, since no normal emails are mistakenly filtered out. However, the under-block error rate of a slack filter is higher than that of a stick filter. The under-block error rate falls from 20.89 to 7.50%, if a stick rather than a slack filter is applied. That is, 92.50% of junk emails would be blocked. Meanwhile, the over-block error rate would rise from 0 to 10.28%. Whether to adopt a slack filter or a stick filter is up to the user.

6. Discussion

Spam is one of the most important problems on email, and is likely to become worsen in the near future. Many studies are aimed at controlling spam. Existing techniques have been shown not to filter out all spam emails. Additional effort is necessary to enhance current anti-spam techniques.

Most current anti-spamming techniques filter junk emails out based on email subjects and body messages. We note that subjects and email contents are not the only cues for spamming judgment. This investigation presents a statistical analysis of the header session messages of junk and normal emails. The statistical analysis results indicate that header session messages could be treated as cues for anti-spamming propose. Content analysis results demonstrate that the message-ID, mail user agent, sender and receiver addresses in the header session are cues for anti-spam filtering, and filtered out up to 92.5% of the junk emails collected in this study.

Some filtering technique may cause a high rate on over-blocking normal emails. Over-blocking normal emails may induce loss of normal emails, and lower the reliability of email services. However, the proposed approach achieves an over-block error rate of zero if the slack filter is adopted. This characteristic of zero over-block error rate is a significant advantage of the proposed anti-spamming approach.

Table 2 – Anti-spamming results

		Actually					
		Normal emails		Solicited listserv and commercial emails		Spam emails	
Judgment	Normal	515	89.72%	57	8.74%	752	7.50%
	Indeterminate	**59**	**10.28%**	**18**	**2.76%**	**1342**	**13.39%**
	Spam	0	0.00%	**577**	**88.50%**	7930	79.11%

Table 3 – Anti-spam efficiency – over and under-block errors

	Slack filter	Stick filter
	Filter out emails judged as spam (%)	Filter out email judged as spam and indeterminate (%)
Over-block errors rate = blocked normal emails/ (unblocked normal emails + blocked normal emails)	0.00	10.28
Under-block errors rate = unblocked junk emails/ (blocked junk emails + unblocked junk emails)	20.89	7.50

The filter out efficiency of the proposed anti-spamming approach can be argued to be insufficiently high. Other proposed methods have filter out rates as high as 98.54% (Soonthornphisaj et al., 2002) or higher. However, the anti-spamming approach proposed herein is a supplementary method, rather than a replacement for other filter out techniques. This proposed concept of employing header session messages to filter out junk emails may coexist with other anti-spamming approaches. No conflict would be found between the proposed idea and existing anti-spamming approaches.

This study conducted a statistical analysis on the content of 10,024 junk emails and 599 emails and 635 solicited listserv or commercial emails contributed by volunteers. Notably, the number of the junk emails examined in the statistical analysis was small. Future studies may undertake a large-scale statistical analysis to verify the results of this study. Obtaining a large-scale repository of junk emails from emails service providers, or from junk email collectors such as spamarchive.org or spam.org, is not a different task. However, normal emails are difficult to collect. Some volunteers may agree to donate some rather than all of their normal emails owing to concerns about privacy. This may lead to a bias in the source data. After overcoming this obstacle, future studies may undertake a large-scale statistical analysis to verify the proposed spam filtering method.

Acknowledgment

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC 93-2416-H-305-002.

REFERENCES

Ahmed S, Mithun F. Word stemming to enhance spam filtering. In: Proceedings of first conference on email and anti-spam, Mountain View, CA; July 2004.

Androutsopoulos I, Paliouras G, Karkaletsis V, Sakkis G, Spyropoulos C, Stamatopoulos P. Learning to filter spam e-mail: a comparison of a naive Bayesian and a memorybased approach. In: Proceedings of the workshop on machine learning and textual information access, Lyon, France; September 2000.

Carreras X, Marquez L. Boosting trees for anti-spam email filtering. In: Proceedings of fourth international conference on recent advances in natural language processing, Tzigov Chark, Bulgaria; September 2001.

Cohen W. Learning rules that classify e-mail. In: Proceedings of AAAI spring symposium on machine learning in information access, Stanford, CA; March 1996.

Cournane A, Hunt R. An analysis of the tools used for the generation and prevention of spam. Computers & Security 2004; 23(2):154–66.

Cranor LF, LaMacchia BA. Spam!. Communications of the ACM 1998;41(8):74–83.

Crocker DH. Standard for the format of APRA internet text messages. The Request for Comments (RFC 822), <http://www.rfc.org>; 1982.

Damiani E, di Vimercati SDC, Paraboschi S, Samarati P, Tironi A, Zaniboni L. Spam attacks: p2p to the rescue. In: Proceedings of the 13th international world wide web conference, New York, NY; May 2004.

Denning PJ. ACM president's letter: electronic junk. Communications of the ACM 1982;25(3):163–5.

Drucker H, Wu D, Vladimir VN. Support vector machines for spam categorization. IEEE Transactions on Neural Networks 1999;10(5):1048–54.

Golbeck J, Hendler J. Reputation network analysis for email filtering. In: Proceedings of first conference on email and anti-spam, Mountain View, CA; July 2004.

Gordon R, Hongyuan Z. Exploring support vector machines and random forests for spam detection. In: Proceedings of first conference on email and anti-spam, Mountain View, CA; July 2004.

Hinde S. Spam: the evolution of a nuisance. Computers & Security 2003;22(6):474–8.

Johnson K. Internet email protocols: a developer's guide. Boston, MA: Addison Wesley; 1999.

Jung J, Sit E. An empirical study of spam traffic and the use of DNS black lists. In: Proceedings of fourth ACM SIGCOMM conference on internet measurement, Taormina, Sicily, Italy; October 2004.

Kolcz A, Alspector J. SVM-based filtering of e-mail spam with content-specific misclassification costs. In: Proceedings of the TextDM'01 workshop on text mining-held at the 2001 IEEE international conference on data mining, San Jose, CA; November 2001.

Leiba B, Borenstein N. A multifaceted approach to spam reduction. In: Proceedings of first conference on email and anti-spam, Mountain View, CA; July 2004.

Postel J. On the junk mail problem. The Request for Comments (RFC 706), <http://www.rfc.org>; 1975.

Provost J. Naive-Bayes vs. rule-learning in classification of email. Technical Report AI-TR-99-284. The University of Texas at Austin, Department of Computer Sciences, <http://www.cs.utexas.edu/users/jp/research/publications/>; 1999.

Resnick PJ, Hansen DL, Richardson CR. Calculating error rates for filtering software. Communications of the ACM 2004;47(9): 67–71.

Rigoutsos I, Huynh T. Chung-Kwei: a pattern-discovery-based System for the automatic identification of unsolicited e-mail messages (SPAM). In: Proceedings of first conference on email and anti-spam, Mountain View, CA; July 2004.

Sahami M, Dumais S, Heckerman D, Horvitz E. A Bayesian approach to filtering junk e-mail. In: Proceedings of AAAI workshop on learning for text categorization, Madison, Wisconsin; July 1998.

Schneider K. A comparison of event models for naive bayes anti-spam e-mail filtering. In: Proceedings of the 11th conference of the European chapter of the association for computational linguistics (EACL'03), Budapest, Hungary; April 2003.

Shih DH, Hsu TE, Lin B. Collaborative spam filtering on multiagent system. In: Proceedings of ACME international conference on Pacific rim management, Chicago, IL; 2004.

Sinclair S. Adapting Bayesian statistical spam filters to the server side. Journal of Computing Sciences in Colleges 2004; 19(5):344–6.

Soonthornphisaj N, Chaikulseriwat K, Tang-On P. Anti-spam filtering: a centroid-based classification approach. In: Proceedings of 2002 international conference on signal proceeding, Beijing, China; August 2002.

Wang C-C. Sender and receiver addresses as cues for anti-spam filtering. Journal of Research and Practice in Information Technology 2004;36(1):3–7.

Woitaszek M, Shaaban M, Czernikowski R, Identifying junk electronic mail in Microsoft Outlook with a support vector machine. In: Proceedings of 2003 symposium on applications and the internet, Orlando, FL; January 2003.

Chih-Chien Wang is currently Associate Professor of Information Management at the National Taipei University, Taiwan. His research interests include anti-spamming, email rumors, Internet addiction, and the impact of Internet on society.

Sheng-Yi Chen earned his MBA degree from Graduate Institute of Information Management of National Taipei University, Taiwan. His research interests include anti-spamming and blog behaviors.