Using Header Session Messages to Filter-out Junk E-mails

Chih-Chien Wang, Sheng-Yi Chen
Graduate Institute of Information Management
National Taipei University

Abstract: Due to the popularity of Internet, e-mail use is the major activity when surfing Internet. However, in recent years, spam has become a major problem that is bothering the use of the e-mail. Many anti-spam filtering techniques have been implemented so far, such as RIPPER rule learning algorithm, Naïve Bayesian classifier, Support Vector Machine, Centroid Based, Decision trees or Memory-base filter. Most existed anti-spamming techniques filter junk e-mails out according to e-mail subjects and body messages. Nevertheless, subjects and e-mail contents are not the only cues for spamming judgment. In this paper, we present a new idea of filtering junk e-mail by utilizing the header session messages. In message head session, besides sender's mail address, receiver's mail address and time etc, users are not interested in other information. This paper conducted two content analyses. The first content analysis adopted 10,024 Junk e-mails collected by Spam Archive (http://spamarchive.org) in a two-months period. The second content analysis adopted 3,482 emails contributed by three volunteers for a one week period. According to content analysis results, this result shows that at most 92.5% of junk e-mails would be filtered out using message-ID, mail user agent, sender and receiver addresses in the header session as cues. In addition, the idea this study proposed may induce zero over block errors rate. This characteristic of zero over block errors rate is an important advantage for the anti-spamming approach this study proposed. This proposed idea of using header session messages to filter-out junk e-mails may coexist with other anti-spamming approaches. Therefore, no conflict would be found between the proposed idea and existing anti-spamming approaches.

Keywords: Web Intelligence and Web Based Information Technology, Spam, Unsolicited E-mail, Junk E-mail, E-mail Address, Filter

I. Introduction

Internet grew vigorously fast and is now a necessary of daily life for many people. According to eTForecasts's statistics, the number of Internet users in the world will surpass 1 billion in mid 2005 and the U.S. continues to lead with over 185 mill Internet users for year-end 2004[11]. According to report conducted by Center for the Digital Future at University of South California Annenberg School, the time American Internet users spend on Internet a week grew from 9.4 hours in 2000 to 12.5 hours in 2003[5]. That is to say, surprisingly, American Internet users spend on Internet up to 1.78 hours everyday. In the year 2000 to 2003, the growth rate of time spending on Internet is 33%. Therefore, Internet has become an important role in people's life.

The survey conduct by Center for the Digital Future at University of South California Annenberg School also indicated that the Internet activity American Internet users the most do is E-mail and instant message, about 90.4% [5]. So, using E-mail is the major activity when surfing Internet. However, spam problem critically influences every Internet users' life and wastes huge resources include network bandwidth and disk storage. Obviously, existence of junk e-mails is an irritating problem when using Internet. Nine out of ten e-mails in America are spam and 76% of all e-mails globally are spam, according to Sharon Gaudin (2004) [26] [27]. The spam problem is going to get worse if we do not put effort on anti-spamming.

Jon Postel, an Internet pioneer, recognized the potential of junk e-mail problem as long ago as November 1975 and proposed Requests for Comments (RFC) 706 that draw up the problem for junk e-mail [17] . In 1982, ACM president Peter J. Denning [21] published the first article about junk e-mail on Communications of ACM. He indicated that junk e-mail will abound in E-mail mailbox. However, "Spam" word became wide-spreading in April 1994, two American lawyers named Canter and Siegel hired an 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 service [19]. After this event, some people identified it as "Spam" and the word caught on. This event can be called the beginning of spam.

Many studies aimed at spamming, for example, Damiani et al. presented a peer-to-peer architecture between mail servers to collaboratively share knowledge about anti-spamming [10]. Jung and Sit focused on DNS black lists [14]. Sahami et al. applied Bayesian approach to filtering junk e-mails out [20]. Rigoutsos and Huynh presented a Chung-Kwei algorithm based on Genetic algorithm to implement a system for the analysis of junk e-mail [13]. Leiba and Borenstein found no one technique solves the spam problem fully, but different techniques excel in different ways and he present using multiple techniques in several layers in order to promote filtering effect [3]. Golbeck and Hendler presented an e-mail scoring mechanism based on a social network augmented with reputation ratings [15]. Goodman focused on the sender's IP address that cannot be faked, so it is a key for anti-spam [7].

The existed techniques could not filter out all spam emails. Spammers might find way to avoid filter out. Junk e-mails may contain faked sender's e-mail address, send in

Proceedings of the Fifth International Conference on Electronic Business, Hong Kong, December 5-9, 2005, pp. 952 - 961.

small batch, and with subjects irrelevant with mail contents to avoid being filtered by anti- spamming mechanisms. Therefore, more efforts are need to improve current anti-spam techniques.

In addition, some filtering technique may result a high rate on filtering spam out, but simultaneously cause a high rate on over blocking normal e-mails. For example, some anti-spamming mechanism might filter out e-mails containing the word “adult” although these solicited e-mails are for “adult continue education” and are sent to or by a scholar majoring in this field. Moreover, a manager might want to e-mail a message to all employees, such bulk e-mails might also be filtered by anti-spam mechanism. At these two situations, the anti-spamming mechanisms have a high over-blocking rate.

Over-blocking normal e-mails as spamming may induce lose of emails and make email service un-reliable. The over-blocking problem may cause users trouble for their daily life or work. Internet users who encounter over-blocking problem may not trust e-mail anymore. People sending out an e-mail have to confirm receiver the delivery of email. This over-blocking problem makes email an inconvenience application.

Most existed anti-spamming techniques are memory-base approaches filtering junk e-mails out according to email subjects and body messages. However, subjects and e-mail contents are not the only cues for spamming judgment. All e-mails have a header session composed of several messages about sender's mail addresses, receiver's mail addresses, mail servers, client e-mail software, message identity number, time stamp, etc. These messages existed in header session may be used as cues for anti-spamming.

Due to the imperfect of existed anti-spamming techniques, more efforts are need to find new anti-spamming approaches. This paper presents a new idea of filtering junk e-mail by utilizing the header session messages. Besides content and subject, e-mail has a header session containing some fields of messages. Past anti-spamming techniques usually used email subjects or contents as cues for anti-spam filtering and neglected the information in the header session.

Header session is designed for storing messages about e-mail delivery. Usually Internet users do not care messages in the header session expect sender e-mail address for replying e-mail and time stamp for sorting e-mails. Most users are not interested in other information. However, these messages which users are not interested in may be cues for anti-spam filtering.

This study use message-ID, mail user agent, sender and receiver addresses in the header session as cues for anti-spam filtering. The e-mail addresses of sender and receiver identify sender and receiver. Spammers usually send junk e-mails with invalid sender address to avoid possible accusation and suspension of e-mail service by Internet Service Provider. In addition, most junk e-mails are sent out in bulk. Spammers usually put receivers’ emails in the field of blind carbon copy and do not reveal the real receivers to avoid receivers to know that this e-mail is send out for enormous copies. That is, the receivers’ e-mail addresses would not been found in receiver and carbon copy receiver field for most junk e-mail.

X-Mailer field of header session indicates what client software or mail user agent (MUA) was used to send the e-mail out. Normal personal communication e-mails are sent out via client software such as outlook, outlook express, lotus note, and so on. However, junk e-mails are sent out via bulk email software. These bulk email software may not point out their software name in the X-Mailer field or randomly put unmeaning characters in the X-mailer field.

Message-ID field is generated either by the MUA or by the first mail transfer agent (MTA) the message passes through, uniquely identifying a piece of e-mail. Most spammer would fake the part value of Message-ID field and cause the domain name of sender address not match the domain part of message-ID. Spammers hope to avoid revealing the real domain name of the MUA or the first MTA the message passes through and thus add a fake domain name to Message-ID field.

The rest of this paper is organized as follows. In the following sections, this paper introduces reviews on anti-spamming mechanisms and efficiency of anti-spam techniques. Then, research design is presented and the results of content analysis are detailed. Finally, we discuss the possibility of using the header session messages as cues for anti-spam filtering. The results of this study indicated that these header session messages might be useful in screening out junk e-mails.

II. Anti-Spamming Approach

The rapid increase in spam traffic took a bothering problem to end users and business corporations. However, a number of methods have been proposed to filter spam out. However, these anti-spamming methods had a very limited effect so far. Various techniques for filtering spam are listed below.

II. 1 Bulk E-mail Filtering

Bulk e-mail filtering is the easiest filtering technique that filtering bulk e-mail out for mail server. This method is based on the assumption that junk e-mail are generally sent to a large number of recipients, so system administrator only need to set a upper limit for recipients of every e-mail at mail server. However, spammer can easily avoid this filtering technique by changing the e-mail header information constantly after sending a certain number of e-mails. In addition, this method may easily filter out important e-mail that includes a large number of recipients. The key determinant of this method’s efficiency is recipients’ upper limit setting. If setting is too strict, this filtering technique would easily filter out the normal e-mails. On the contrary, setting too loose would make junk e-mail not be filtered out.

II. 2 Filtering by Keyword

Keyword filtering is the most frequent used anti-spamming

technique. There are two approaches for this method. The first one is an easy but inefficient approach which sets keyword and filter out all e-mails with keyword appeared in subject or content. “On Sale”, “Sex” and “Get Rich” are frequent used keyword. However, this method is inefficiency, while spammer would avoid using these keywords and normal e-mails containing these keywords will be mistakenly regarded as junk e-mails.

The second way is to filter out junk e-mails basing on machine learning results. This approach is more complex and more efficient than the former one. In this approach, the keywords are determined by machine learning algorithm and frequency of all keyword is calculated to discriminate junk and normal e-mails. Many studies are based on this approach, such as RIPPER rule learning algorithm (Cohen, 1996[7]; Provost, 1999[22]), Naïve Bayesian classifier (Sahami 1998[24]; Schneider 2003[25]; Sinclair 2004[29]), Support Vector Machine (Drucker et al. 1999[9]; Kolcz and Alspector 2001[18]; Gordon and Hongyuan 2004[12]), Centroid Based (Soonthornphisaj et al. 2002[30]), Decision trees (Carreras and Marquez 2001[4]) or Memory-base filter (Androutsopoulos et al. 2000[2]). This method is based on the assumption that the subject or body sector of junk e-mails may contain specific words. However, this technique does not work perfectly, and thus normal e-mails would be filtered out simply because they include too many words on the keyword list for junk e-mails.

II. 3 Black List

This approach exploits a black list database to block specific address or domain name of e-mail. Such database is made available on the Internet, such as DNSBLs (http://dsbl.org /main) or SORBs (http://www.us.sorbs.net) . If the database updates frequently, it can be reliable for filtering certain known spammers’ address out. However, it is a defect that if mail servers improperly set normal e-mail address or domain name into black lists would filter out normal or important e-mail. Besides, spammers usually leave random assigned faked sender addresses which not in the black list. Black list approach can not function if sender addresses are faked.

II. 4 White List

White list is design to avoid filtering normal e-mails out. White lists gather permitted e-mail address or domain name and often collaborate with black lists. Black lists block illegal e-mails address and White lists allow normal e-mails to pass. It is more difficult to maintain the white lists database perfectly than black list.

II. 5 Sender Address Validity

In 1982, Crocker revised the Requests for Comments (RFC) 822 that is a standard for the format of ARPA Internet text messages and the format of e-mail is described in. [8] According to RFC 822, each e-mail must include the field of “From” that contains the address of the sender who wished this message to be sent. All e-mails should have at least one sender and the "From" field must be present. Although RFC 822 mandate the existence of the “From” header in e-mails, the sender address can be invalid or faked by spammer. The spammers avoid being accused of sending junk e-mail and breaking the law. Additionally, most Internet Service Providers (ISPs) or e-mail service provides would suspend the use of spammers’ e-mail address or refuse to relay the e-mails they sent. According to Wang (2004)’s research, 60.3% junk e-mails provide invalid sender address [31]. Therefore, the validity of sender address left in the “From” header session may be a cue for anti-spam filtering.

II. 6 Receiver Address as Cue

Also according to RFC 822 [8], the receiver addresses are list in the “TO”, “CC”, or “BCC” fields and like sender address, each e-mail must have at least one receiver. The “TO”, “CC”, or “BCC” headers are used to present the recipients of this e-mail where to sent. “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 are very similar. However, “BCC” differs from “TO” and “CC”. “BCC” standing 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. According to Wang’s research, only 7.2% spam would put receiver address in “TO” or “CC” and spammer usually use the “BCC” for the receiver address to avoid revealing that junk e-mail are sent in bulk.[31] Hence, the presence of receiver address left in the “TO” or “CC” header session may be a cue for anti-spam filtering. However, this approach may filter out normal e-mail while people use the “BCC” for the receiver address to send normal e-mails.

II. 7 Mail User Agent as Cue

In header session, RFC 822 notes the use of X- at the beginning of field names to indicate that a field is an extension. The X-Mailer field indicates what e-mail client program or MUA was used to generate the e-mail. Although this field is not required in header session, most MUA developers generally have their software add an appropriate X-Mailer field to all out-bound e-mails. According to observation made by this study, most junk e-mail do not include the X-Mailer field or include this field with random assign value. Therefore, the field of X-Mailer may be a cue for anti-spam filtering. Most frequent used MUAs, for example, Outlook Express, MS Outlook, Lotus Note or Eudora etc., marked X-Mailer field of out-bound e-mails exactly. On the contrary, an inbound e-mail that X-Mailer field value is null, meaningless random assign value may mean that this e-mail is probable junk e-mail. A normal MUA for sending bulk e-mail 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 although developing bulk e-mail software do not violate any law. As a result, the X-Mailer field can be a cue for ant-spam filtering.

II. 8 Message-ID as Cue

The unique message identifier in the header session is generated by the MUA or by the first MTA the message passes through if MUA did not yet assign one for the e-mail. This identifier is intended to be machine readable and not necessarily meaningful to humans. The format of this message ID field value is with a symbol of "@" dividing the value into two parts. The left side contains a string of characters to uniquely identify the message on the machine where it was created and is usually based on the date and time or depending on the e-mail software generating the data. The right side specifies that machine or domain name [16: pp. 32-33]. Most spammer would fake this domain value to avoid possible internet service suspend and cause the domain of sender address not match the domain part of Message-ID. Spammers hope to avoid revealing the real domain name of the MUA or the first MTA the message passes through and thus add a not existed domain name. Consequently, this condition provides a cue for deciding the possibility of an incoming e-mail is a junk-mail.

III. Efficiency of Anti-Spam Techniques

While calculating the effectiveness for anti-spam filtering techniques, two types of errors should be taken into consider. First, under-blocking occurs when e-mail is not blocked that should be filtered out. Second, over-blocking occurs when solicited normal e-mail that should not be filtered out is blocked. Shortly, it is bad anti-spamming techniques that junk e-mails are not blocked or normal e-mails are blocked. These two error rate format proposed by Resnick et al. (2004) are listed blow [23].

Under-blocking errors = unblocked junk e-mails / (blocked junk e-mails + unblocked junk e-mails)

Over-blocking errors = blocked normal e-mails / (unblocked normal e-mails + blocked normal e-mails)

Reducing both two error rates mentioning above is a good filtering technique should do. However, the importance of under-blocking errors and over-blocking errors is not same. For most e-mail users, the problem of over-blocking errors is more important than under-blocking errors. While encountering unblocked junk e-mails, users only spend additional time on deleting them. However, over-blocking normal e-mails generally can not be recovered. Thus, users would lose some important messages and it may cause troubles at communication for work or daily life. If e-mail users aware the possible risk of over-blocking, they have to ask receiver to confirm e-mail receiving. This may bring inconvenience to e-mail users and make e-mails unreliable.

The under-blocking errors and over-blocking errors are benchmark for effectiveness of anti-spam filtering techniques. However, most past researches focus on the under-blocking errors although it is important to reduce both two error rates for anti-spam filtering. For examples, Chen et al. [6] report that they can filter out 98.54% junk e-mail but do not mention the over-blocking error rate. Woitaszek et al. [32] report as low as 96.69% under-blocking errors but not mentioning the over-blocking error. The same situations had also found in Ahmed et al.[1], Androutsopoulos et al. [2], Drucker et al. [9], Shih et al. [28] and Soonthornphisaj et al. [30]'s studies.

IV. Content Analysis for Junk and Normal E-Mails

This study conducted two content analyses. The first content analysis (study 1) adopted 10,024 Junk e-mails collected by Spam Archive (http://spamarchive.org) in a two-months period. Spam Archive has collected large number junk e-mails that is donated by end users and is a well known large spam repository for developing anti-spam tools. The second content analysis (study 2) adopted 3,482 emails contributed by three volunteers for a one week period. The collected 3,482 emails in study 2 were classified into three categories: normal, junk and solicited listserv and commercial ones. Normal e-mails are e-mail for personal communication. Junk e-mails are unsolicited e-mails and are usually send in bulk and for commercial purpose. However, some emails are solicited for users although they are sent in bulk. Listserv e-mail is a typical case for this category. People may join a listserv, discussion board, or family in yahoo to receive e-mails. In addition, users may subscribe retailing websites to receive updated sale messages. These kinds of listserv and commercial e-mails should be treated as solicited although most of them are sent in bulk.

This study use content analysis to examine the possibility of using header session messages as cues to discriminate normal and junk e-mails. Analyzing fields of sender, receiver addresses, messages ID, and MUA in normal and junk e-mails' header sessions are used for this purpose.

IV. 1 Sender Addresses

Sender addresses validity was check for all normal and junk e-mails this study collected. The study checked sender addresses via Domain Name Server (DNS) checking for existence of mail server and Simple Mail Transfer Protocol (SMTP) checking for existence of mail account. Each sender' email account was check if the SMPT servers refused to response to email addresses validity checks. Figure 1 and 2 indicates the sender address checking results of study 1 and 2. Of the 10,024 junk e-mails study 1 collected, 6,664 (66.48%) were with invalid sender addresses. Of the 2,248 e-mails study 2 collected, 791 (35.1%) were with invalid sender addresses. Of the 635 solicited listserv or commercial e-mails study 2 collect, only 28 (4.41%) were with invalid sender addresses. Moreover, none (0%) of normal e-mail study 2 collected was with invalid address.

The results of sender address validity checking showed that sender address may be a cue for reduce over block rate.

As figure 2 indicated, all normal e-mails were with valid sender addresses. This means that there is no side effect, no normal email will be mistakenly block out, if we filter out all e-mails with invalid sender addresses. Simply filter out e-mail without valid sender addresses may block out 66.48% junk e-mails in study 1 and 35.1% junk emails in study 2. However, 4.41% solicited listserv or commercial e-mails in study 2 will also be filter out according to this valid sender address rule. This 4.41% filtered emails might be regarded as over block e-mails if we treat solicited listserv and commercial ones as normal. Nevertheless, the over block rate will be zero if users think that it is acceptable to filter out listserv and commercial e-mails.

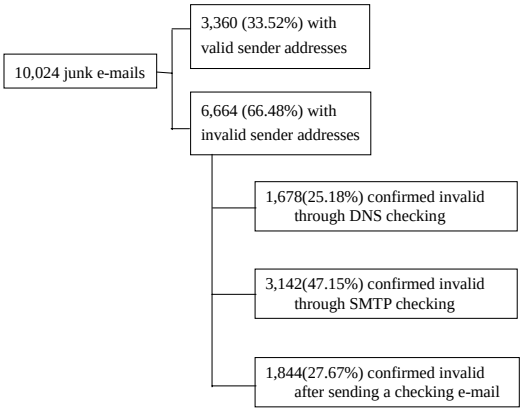


Figure 1: Sender Address Checking for Study 1 (10,024 Junk e-mails)

IV. 2 Receiver Addresses

Presence of receiver address left in the “TO” or “CC” header session may be a cue for anti-spam filtering since that spammer usually use the “BCC” for the receiver address to avoid revealing that junk e-mail are sent in bulk. [31]

All receiver addresses of junk e-mails collected by spamarchive.org are omitted. So this study cannot analyze the receiver addresses of junk e-mail in study 1. The figure 3 indicated the content analysis results of receiver addresses in study 2.

It is showed that 84.64% normal e-mails containing receivers’ address in “TO” and “CC” fields while 15.36% normal e-mails without receivers’ address in “TO” and “CC” fields. In addition, 44.26% junk e-mails and 11.81% solicited listserv and commercial e-mails were with receivers’ address in “TO” and “CC” fields while 55.74% junk e-mails and 88.19% solicited listserv and commercial e-mails without.

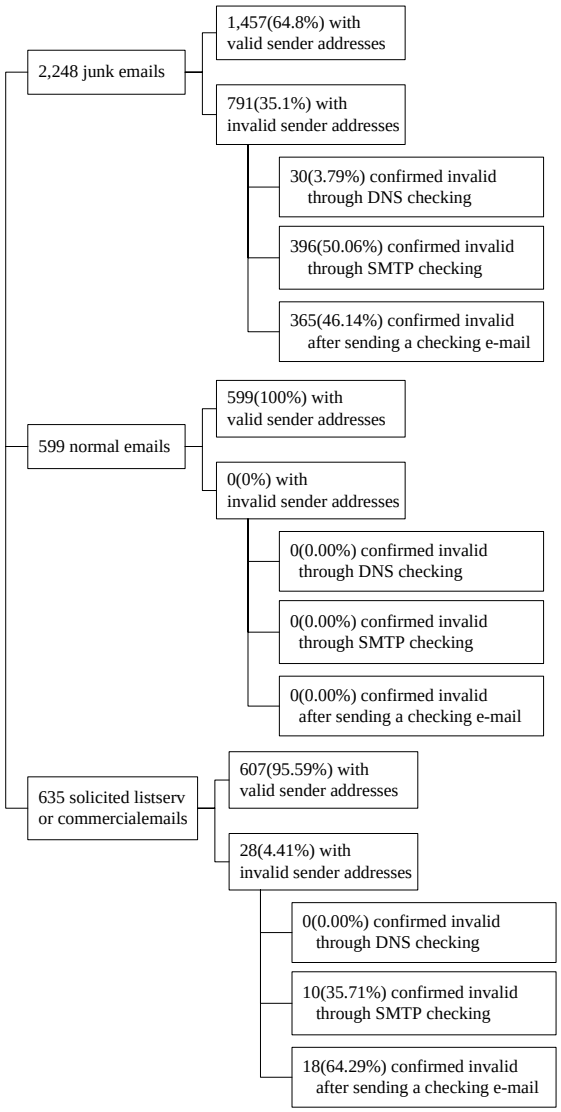


Figure 2: Sender Address Checking for Study 2 (3,482 normal, junk and solicited listserv and commercial e-mails)

Receiver address could be used as a cue for normal e-mail judgment. If an e-mail is with receivers’ address in “TO” and “CC” fields, the possibility of this e-mail is normal e-mail is high. However, this should be an assist cue since that some normal e-mails purposely put receiver’ e-mail addresses in the “BCC” header session.

The high block out percentage for solicited listserv and commercial e-mails are from the fact that most listserv and commercial e-mails are sent in bulk. It stands to reason to put receiver’s addresses in “BCC” rather than “TO” or “CC” fields if e-mails are sent in bulk. Over block some normal e-mails (15.36%) and most solicited listserv and commercial e-mails (88.19%) are side effect of using receiver addresses to filter out e-mails. However, since that most normal emails were with receiver addresses in “TO” or “CC” fields, receiver addresses may still be assist cues for normal e-mails. In addition, receiver addresses may still be useful in anti-spamming if it is acceptable for users that over-blocking

solicited listserv and commercial e-mails.

IV. 3 Mail User Agent

Mail User Agents (MUAs) are e-mail client programs used to generate the e-mail. Most normal e-mails were with X-mailer messages to present the MUA which sent the e-mails although this is not required. However, most junk e-mails do not include the X-Mailer field or include this field with random assign value.

Figure 4 and 5 indicated the content analysis results of X-mailer field. The results showed that 1.73% junk e-mails in study 1 and 4.34% junk e-mails in study 2 were sent by frequent used bulk e-mail programs. In addition, no X-mailer messages found in 58.21% junk e-mails in study 1 and 55.05% in study 2, respectively. 5.69% junk e-mail in study 1 and 9.75% in study 2 were with random assigned value for X-mailer field. 2.03% junk e-mail in study 1 and 2.75% in study were sent by infrequent used MUA.

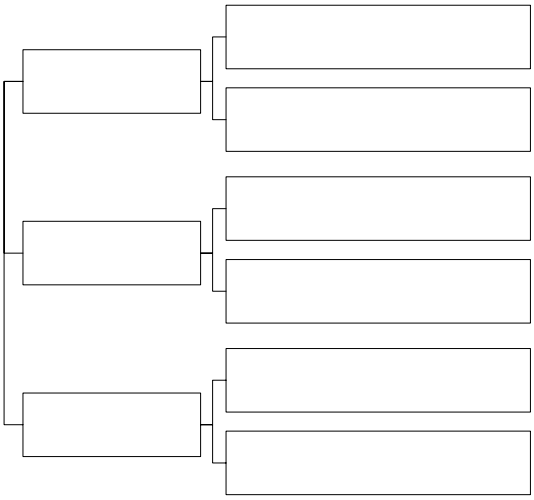


Figure 3: Receiver Addresses of Study 2

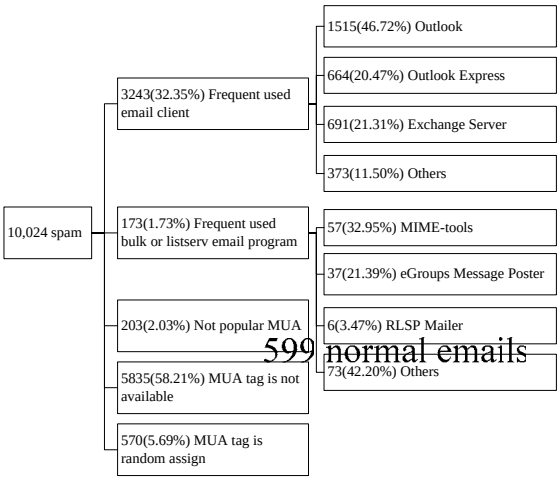


Figure 4: MUA for Study 1 (10,024 Junk e-mails)

Moreover, most normal e-mails (52.96%) in study 2 were sent by frequent used MUA program as figure 5 indicated. Only 2.44% normal e-mails in study 2 were sent by bulk e-mails program, 0.87% by infrequent used MUA, 0.00% with random assign X-mailer value, and 43.73% were without X-mailer message.

As content analysis results point out, MUA could be used as a cue for normal e-mail judgment. The possibility of the e-mail is normal one is high if an e-mail is with frequent used MUA. On the contrary, the possibility of the e-mail is junk one is high if e-mails are sent by bulk e-mail program or the X-mailer messages are random assigned. However, this should be an assist cue since that normal e-mails send by web mail interface may have similar characteristics of X-mailer field.

IV. 4 Message-ID

Message-ID is generated by the MUA or by the first MTA the message passes through if MUA did not yet assign one for the e-mail. This may be used as a judgment cue for normal e-mail. Most spammers hope to avoid revealing the real domain name of the MUA or the first MTA the message passes through and thus add a not existed domain name. So, if the sender addresses match the domain name specific in message-ID, the possibility is high of the e-mails are normal ones. However, it is not definitely junk email if the message-ID does not match sender address since that message-ID may be assign by MUA rather than MTA. If message-ID is assign by MUA such as outlook express, the computer name rather than e-mail domain name is put in right side of message-ID. Besides, according to the authors' observation, some web mail system put software or computer name rather than domain name to right side of message-ID. This also makes the not-match between message-ID and sender address.

Figure 6 and 7 indicated the analyze results of match between message-ID and sender Address. The results showed that message-IDs did not match sender address for 82.66% junk e-mails in study 1 and 39.32% junk e-mails in study 2. However, only 11.02% message-IDs of normal e-mails in study did not match sender addresses.

V. Using Header Session Message to Anti-Spam

As mention above, sender address, receiver address, MUA, and message-ID could be used as cues for anti-spamming. If sender address is invalid, the e-mail can not be normal according to content analysis results, as figures 1 and 2 indicated. If the receiver address is not found in "TO" or "CC" fields, the possibility is high of the e-mail is junk, as figure 3 pointed out. Besides, as figures 4 and 5 indicated, the possibility is high of an e-mail is send in bulk if MUA is bulk email program, MUA tag is not available, or is random assign. In this situation, the e-mail may be junk and should be filter out. Moreover, if message-ID does not match sender address, the possibility is high that the message-ID is

507(84.64%) contain receiver's address in "TO" and "CC" fields

599 normal emails

92(15.36%) without receiver's address in "TO" and "CC" fields

995(44.26%) contain receiver's address in "TO" and "CC" fields

2,248 junk emails

1253(55.74%) without receiver's address in "TO" and "CC" fields

assigned by MUA rather than MTA or message-ID is faked to avoid possible trace. Some frequent used MUA programs, such as outlook express and some webmail programs, assign Message-IDs to all send-out e-mails base on their own rules irrelative to sender's address so that the Message-ID would not match senders' address. However, if an e-mail is not send by these kinds of MUA, the message-ID should match sender's address. If not match, the possibility is high of the e-mail is junk, as this study found in figure 6 and 7.

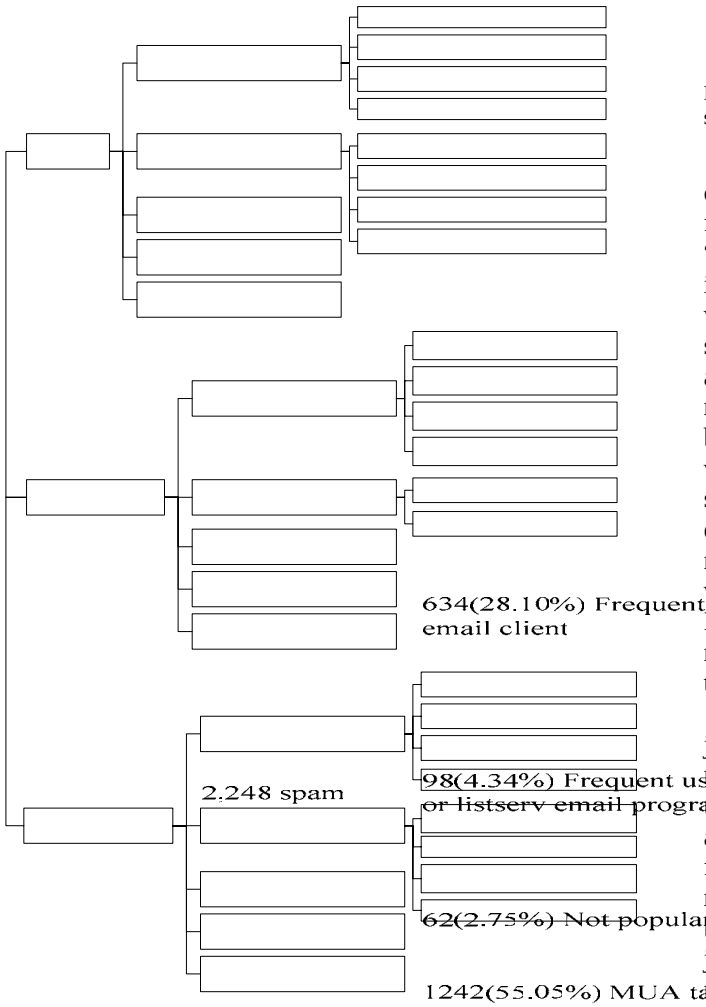


Figure 5: MUA for Study 2 (3,482 normal, junk and solicited listserv and commercial e-mails)

Figure 6 Match between Message ID and Sender Address for junk e-mails

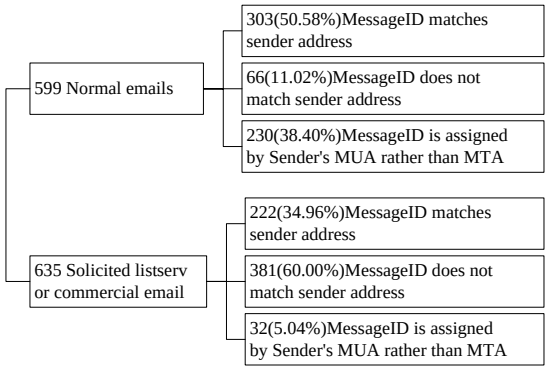


Figure 7 Match between Message ID and Sender Address for normal and solicited e-mails

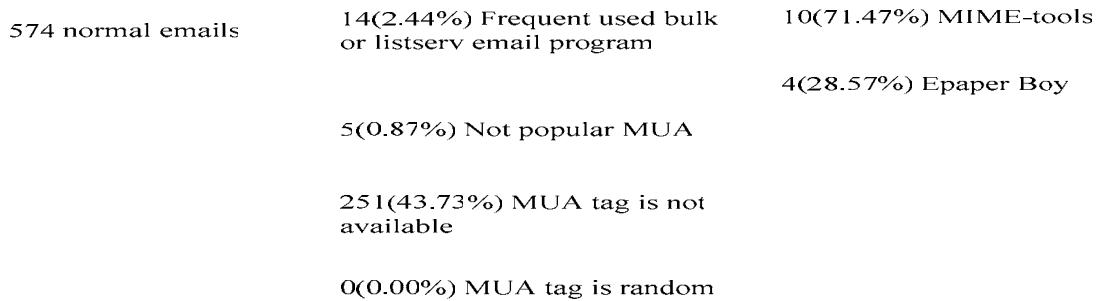
Sender address, receiver address, MUA, and message-ID can not to be used alone for anti-spamming. If we simply filter out an e-mail without receiver address in "TO" and "CC" fields, we may over block some normal e-mails which is sent purposely by blind carbon copy. The similar situation will also be found in using MUA and message-ID as anti-spamming cues. Image the situation that a user hopes to send a personal notify to her or his friends in a large batch, he may adopt bulk e-mail MUA. Filter out e-mails send out by bulk e-mail MUA will also over block this kind of e-mails which should be treat as normal rather than spam. The same situation will happen for message-ID. It still could be normal e-mail when message-ID does not match sender' domain name. Some SMTP components, modules and libraries for website programming do not consider well about message-ID tag. For those users who use these kinds of MUA, message-IDs would not match senders' addresses, even what these users send out are normal emails.

This study proposed a combined anti-spamming judgment approach since single item of sender address, receiver address, MUA, and message-ID can not be used as the only cue for anti-spamming. This anti-spamming approach judge e-mails as normal or spamming according to four cues, i.e. sender address, receiver address, MUA, and message-ID. An e-mail will be judge as normal if it conform the rules of normal e-mail. Nevertheless, it will be judge as junk when conform the rules of junk. If an e-mail conform neither junk nor normal, the e-mail will be classified as indeterminate. Filter or not this kind e-mail should be user's personal choice.

Table 1 indicated the anti-spamming approach this study proposed. The possibility is high of an e-mail is not spam if it is with valid sender address, MUA is not frequent used bulk or automated email program, and Message-ID matches sender address or is assigned by MUA rather than sender's MTA.

However, if an e-mail is probably sent by bulk e-mail program, Message-ID does not match sender address, is carbon copy to receivers, and is with invalid sender addresses, the possibility of the e-mail is junk.

Qualifications for spam rules mentioned in table 1 are too strict. An e-mail could still be spam if it does not reach



all four rules mention in table 1. This study proposed an idea that an e-mail will be regards as spamming if with invalid sender address or if it matches two of remain three rules, rules 2, 3 and 4, of table 1. This study advocated that rule 1, invalid sender address, is enough to filter out an e-mail although some listserv e-mails will be mistakenly filter out. It is unreason to send an e-mail with valid sender address even it is sent automatically.

Table 2 and 3 indicated the anti-spamming results of study 1 and 2. The junk e-mail filter out rate for study 1 is 79.11% if we filter out the e-mails which are judged as spam. However, this rate will raise 13.39% to 92.50% if filtering out e-mails which are judged as indeterminate and keep only e-mails which are judged as normal. For study 2, the junk e-mail filter out rate is 75.66% if we filter out the e-mails which are judged as spam. This filter out rate will raise 2.97% to 78.63% if filtering out e-mails which are judged as indeterminate. However, the anti-spamming approach this study proposed filtered out 88.50% solicited listserv and commercial e-mails. This rate would increase 2.76% to 91.26% if filter out e-mails which are judged as indeterminate. Most solicited listserv and commercial e-mails look like junk e-mails. The proposed anti-spamming approach would filter most of them, 88.50%, as table 3 indicated.

Table 4 summary anti-spam efficiency of the approach this study proposed. If a user who hope to filter out junk e-mails as many as possible, he can choice to keep only e-mails which are judged as normal and filter out both spam and indeterminate. This may be named as stick filter. However, some normal e-mails are judged as indeterminate as table 3 indicated. Users have to accept the risk of mistakenly filtering normal e-mails out. On the contrary, users may choice a safe strategy and filter out only e-mails which are judged as spam. In this situation, all normal and indeterminate e-mails are kept. This may be named as slack filter since that only confirmed spam are block.

The over block errors rate reflect the phenomenon that normal e-mail that should not be filtered out is blocked. It is not available for study 1 since that study 1 containing spam e-mails only. The block errors rate is zero if slack filter is adopted for study 2. This means that there is no side effect if slack filter is adopted. No normal e-mails would be filtered out mistakenly. However, the under block errors rate of slack filter is high when comparing with stick filter. The under block errors rate would reduce from 20.89% to 7.50% in study 1 and 24.34% to 21.37% in study 2, if stick rather than slack filter is adopted. This means that 92.50% junk e-mails in study 1 and 78.63% in study 2 would be blocked. Meanwhile, over block errors rate would increase from zero to 10.28%. It is users' own choice that adopting slack or stick filter.

VI. Discussion

Spam is one of the most important problems which going to get worse. Many studies aimed at spamming. However, the existed techniques could not filter out all spam emails. More efforts are need to improve current anti-spam techniques.

Most existed anti-spamming techniques filter junk e-mails out according to e-mail subjects and body messages. However, subjects and e-mail contents are not the only cues for spamming judgment. This paper presents a new idea of filtering junk e-mail by utilizing the header session messages.

According to content analysis results, this study found that message-ID, mail user agent, sender and receiver addresses in the header session as cues for anti-spam filtering. At most 92.5% of junk e-mails would be filtered out using message-ID, mail user agent, sender and receiver addresses in the header session as cues.

Besides, some filtering technique may cause a high rate on over blocking normal e-mails. Over-blocking normal e-mails may induce lose of normal emails and make email service un-reliable. However, the idea this study proposed may induce zero over block errors rate if slack filter is adopt. This characteristic of zero over block errors rate is an important advantage for the anti-spamming approach this study proposed.

Some may argue that the filter out efficiency is not too high of anti-spamming approach this study proposed. Some studies may have filter out rate of as high as 98.54% [30] or 96.69% [31]. However, the anti-spamming approach this study proposed is a supplementary rather than a replacement for other filter out techniques. This proposed idea of using header session messages to filter-out junk e-mails may coexist with other anti-spamming approaches. No conflict would be found between the proposed idea and existing anti-spamming approaches.

Acknowledgment

The authors would like to thank the Taiwan National Science Council for financially supporting this research under Contract No. NSC93-2416-H-305-002.

References

[1] Ahmed, S. & Mithun, F. "Word Stemming to Enhance Spam Filtering, " *The First Conference on Email and Anti-Spam*, 2004

[2] 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, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD 2000)*, 2000.

[3] Barry Leiba & Nathaniel Borenstein. "A Multifaceted Approach to Spam Reduction," *The First Conference on Email and Anti-Spam*, 2004

[4] Carreras, X. & Marquez, L. "Boosting Trees for Anti-Spam Email Filtering," *Proceedings of {RANLP}-01, 4th International Conference on Recent Advances in Natural Language Processing*, 2001.

[5] Center for the Digital Future at University of South California Annenberg School, Digital Future Report- Year4 -2004. http://www.digitalcenter.org/pages/current_report.asp ?intGlobalId= 19

[6] Chen D., Chen T., & Ming H., "Spam Email Filter Using Naïve Bayesian, Decision Tree, Neural Network, and AdaBoost," http://www.cs.iastate.edu/-ton~ie/spamfilter/paper.pdf
[7] Cohen, W. "Learning rules that classify e-mail," *Spring Symposium on Machine Learning in Information Access*, 1996.
[8] Crocker, D. H. 1982, "Standard for the format of APRA Internet text messages," *The Request for Comments (RFC) 822*, http://www.rfc.org.
[9] Drucker, H., Wu Donghui & V. N. Vladimir. "Support Vector Machines for Spam Categorization," *IEEE Transactions on Neural Networks*, 1999.
[10] Ernesto Damiani, Sabrina De Capitani di Vimercati, Stefano Paraboschi, Pierangela Samarati, Andrea Tironi & Luca Zaniboni. "Spam attacks: p2p to the rescue," *Proceedings of the 13th international World Wide Web conference on Alternate track papers*, 2004
[11] eTForecasts**,** Worldwide Internet Users will Top 1 Billion in 2005 http://www.etforecasts.com/pr/pr904.htm
[12] Gordon R. & Hongyuan Z. "Exploring Support Vector Machines and Random Forests for Spam Detection," *CEAS 2004*, 2004.
[13] Isidore Rigoutsos & Tien Huynh. "Chung-Kwei: a Pattern-discovery-based System for the Automatic Identification of Unsolicited E-mail Messages (SPAM)," *The First Conference on Email and Anti-Spam*, 2004
[14] Jaeyeon Jung & Emil Sit. "An empirical study of spam traffic and the use of DNS black lists," *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 2004
[15] Jennifer Golbeck & James Hendler. "Reputation Network Analysis for Email Filtering," *The First Conference on Email and Anti-Spam*, 2004
[16] Johnson, K. Internet Email Protocols: *A Developer's Guide*, Addison Wesley, 1999.
[17] Jon Postel, November 1975, "On the junk mail problem," *the Request for Comments (RFC)*: 706.
[18] 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*, 2001.
[19] Lorrie Faith Cranor and Brian A. LaMacchia, August 1998 , "Spam!," Communications of the ACM, 41 (8), 74-83. The original definition of spam is a luncheon meat canned "Shoulder Pork and hAM" or "SPiced hAM", a product of Hormel Foods (http://www.hormel.com). Most consumers intuitively associate "Spam" with "no nutritive." Some people have indefinite recognition that it came at first from the sketch on the television series "Monty Python's Flying Circus" in England. In this television series, a restaurant server all its meals with lots of spam branded meat and the waitress always repeats the word "spam" several times in describing how much spam branded meat is available in their meals.
[20] Mehran Sahami, Susan Dumais, David Heckerman and Eric Horvitz. "A Bayesian Approach to Filtering Junk E-Mail," *AAAI Technical Report* WS-98-05, 1988
[21] Peter J. Denning , 1982, "ACM president's letter: electronic junk," *Communications of the ACM*, 25(3), 163-165.
[22] Provost, J. "Naive-Bayes vs. Rule-Learning in Classification of Email," *Technical Report*, 1999, 99-284.
[23] Resnick, P. J., Hansen, D. L. & Richardson, C. R. "Calculating error rates for filtering software," *Communications of the ACM*., 2004.
[24] Sahami, M., Dumais, S.,Heckerman, D., & Horvitz, E. "A Bayesian Approach to Filtering Junk E-Mail," *Learning for Text Categorization: Papers from the 1998 Workshop*, 1998.
[25] Schneider, K. "A comparison of event models for naive bayes anti-spam e-mail filtering," *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, 2003.
[26] Sharon Gaudin, Nine out of 10 U.S. Emails Now Spam, June 8, 2004, http://www.esecurityplanet.com/ trends/ article.php/3365341.
[27] Sharon Gaudin, U.S. Sending More Than Half of All Spam, July 1, 2004 http://www.internetnews.com /stats /article.php /3376331.
[28] Shih, D. H., T. E. Hsu & B. Lin. "Collaborative Spam Filtering on Multiagent System," *The XIV ACME Int. Conf. on Pacific Rim Management*, 2004, pp619-624.
[29] Sinclair, S. "Adapting Bayesian statistical spam filters to the server side," *J.Comput.Small Coll*, 2004.
[30] Soonthornphisaj, N., K. Chaikulseriwat, & P. Tang-On. "Anti-Spam Filtering: A Centroid-Based Classification Approach," *Proceedings of 2002 6th International Conference on Signal Proceeding*, 2002.
[31] Wang, Chih-Chien, 2004, "Sender and Receiver Addresses as Cues for Anti-Spam Filtering," *Journal of Research and Practice in Information Technology*, 36 (1), February, pp. 3-7.
[32] Woitaszek M., Shaaban M., & Czernikowski R., "Identifying junk electronic mail in Microsoft Outlook with a support vector machine," *Proceedings of the 2003 Symposium on Applications and the Internet*, 2003, pp.166-169.

Table 1 Anti-Spamming Approach using sender address, receiver address, MUA, and message-ID

Judgment	Approach	Rules
Judged as Normal E-mails	Do not filter out e-mails containing all three characteristic	Normal email has following characteristics. 1. Valid sender address. 2. MUA is not frequent used bulk or automatic e-mail program. 3. Message-ID matches sender address or is assigned by MUA rather than sender's MTA.
Judged as Spam	Filter out e-mails which match spam rule 1. Or match two rules of rules 2, 3, and 4.	Spam has following characteristics. 1. Invalid sender address. 2. E-mail is not send to or carbon copy to receiver. 3. Sender's MUA is bulk email program or MUA tag is not available or random assign. 4. Message-ID does not match sender address.
Judged as Indeterminate	Neither normal or spam e-mails	Neither normal or spam e-mails

Table 2 Anti-Spamming Results for Study 1

		Actually Spam	
Judgment	Normal	752	7.50%
	Indeterminate	**1342**	**13.39%**
	Spam	7930	79.11%

Table 3 Anti-Spamming Results for Study 2

		Actuall y					
		Normal E-mails		Solicited Listserv and Commercial E-mails		Spam E-mails	
Judg ment	Normal	515	89.72%	57	8.74%	482	21.37%
	Indeterminate	**59**	**10.28 %**	**18**	**2.76%**	**67**	**2.97%**
	Spam	0	0.00%	**577**	**88.50 %**	1707	75.66%

Table 4 Anti-Spam Efficiency - Over and Under Block Errors

		Slack filter	Stick filter
		Filter out E-mails judged as spam	Filter out E-mail judge as spam and indeterminate
Over block errors rate = blocked normal e-mails / (unblocked normal e-mails + blocked normal e-mails)	Study 2	0.00%	10.28%
Under Block errors rate = unblocked junk e-mails / (blocked junk e-mails + unblocked junk e-mails)	Study 1	20.89%	7.50%
	Study 2	24.34%	21.37%