Sender and Receiver Addresses as Cues for Anti-Spam Filtering

Chih-Chien Wang

Graduate Institute of Information Management
National Taipei University
69, Sec. 2, JianGuo N. Rd., Taipei City 104-33, Taiwan
wangson@mail.ntpu.edu.tw
http://wangson.idv.tw
Tel: +886-2-2500-9501
Fax: +886-2-2517-0078

This study analysed the sender and receiver addresses of 3,417 unsolicited e-mails. Over 60.3% of unsolicited e-mails were found to have an invalid sender address and 92.8% receiver addresses did not appear in the "To" or "CC" headers. The analytical results indicated that e-mail addresses in the header could provide a cue for filtering junk e-mails.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Information filtering

Keywords: Spam, unsolicited e-mail, junk mail, e-mail address, filter

1. INTRODUCTION

Junk e-mail has long been a problem on the Internet. Internet pioneer Jon Postel recognised the potential for this problem to develop as long ago as November 1975 (Postel, 1975). The problem has now become extremely serious. The growing popularity and low cost of e-mails have attracted the attention of marketers. Using readily available bulk e-mail software and lists of e-mail addresses harvested from web pages and newsgroup archives, sending messages to millions of recipients is very easy and very cheap, and can be considered almost free. Consequently, these unsolicited e-mails bother users and fill their e-mail folders with unwanted messages. Few users, if any, have never received unsolicited e-mails. These unwanted messages generally are called unsolicited e-mails or spam. Spam also describes the action of sending out such mails. These unsolicited e-mails are also known as bulk mails, because they are generally sent out in large batches, and as junk mail, because they are worthless to most recipients.

Spamming interferes with ordinary e-mail communication and reduces employee productivity. Cranor and LaMacchia (1998) reported that spam messages constituted 10% of all incoming messages. It has become more serious recently. It is estimated that there will be more spam than real e-mail and spam currently costs business US $ 13 billion annually (Swartz, 2003). According to Nucleus Research (2003), spamming will cost 1.4 percent of employee productivity, or $874 per year per employee, in 2003. Consequently, screening out spam is an extremely urgent problem.

Manuscript received: 24 July 2003
Communicating Editor: Sidney A. Morris

Several techniques exist for filtering spam. However, none of these filters is perfect. This study proposes a filtering approach based on screening the e-mail addresses in the header section. This study conducted a content analysis of junk e-mail addresses to check the possibility of using e-mail addresses as a cue for filtering spam. The results indicated that invalid sender and irrelative receiver e-mail addresses left in the header section of junk e-mails could be used to screen out junk e-mail. The results of this study may be useful in developing anti-spam strategies.

2. ANTI-SPAM FILTERING TECHNIQUES

Most, if not all, e-mail service providers and Internet users wish to filter out junk e-mails. Various methods of filtering have already been developed. Various methods for filtering spam out are listed below.

2.1 Filtering by Number of Recipients

This method involves screening out e-mails being sent to a large number of recipients. This method is based on the assumption that junk e-mails generally are sent in bulk. However, this method is easily defeated by bulk e-mail programs that automatically reconnect to the mail server or automatically change the e-mail title and/or body messages after sending a certain number of e-mails. The spammer can set this number to as little as twenty or less. Additionally, this method assumes all bulk e-mails to be junk, and thus may filter out important and solicited e-mail as well as unsolicited e-mail.

2.2 Filtering by Keyword

Keyword filtering is another method of screening junk e-mail. Several e-mail service providers currently provide content-based filtering on their mail servers. Some studies based on the Naive Bayes filter (such as Pazzani, 2000; Robinson, 2003) or memory-based filter (such as Sakkis, Androutsopoulos, Paliouras, Karkaletsis, Spyropoulos, and Stamatopoulos, 2003) focus on promoting the filter efficiency of this method.

However, this filter method also does not perform perfectly, and thus users still must check filtered e-mails and put any unsolicited e-mail in the junk mail pool. This method is based on the assumption that the titles and/or body messages of junk e-mails may contain specific words. “Sell”, “on sale”, and “sex” are some typical keywords used to filter e-mails with this approach. These words should be defined in advance. However, this approach suffers the problem of a high possibility of solicited e-mails being filtered out simply because they happen to include words on the filter list.

2.3 Filtering by Sender Address

Filtering spam out based on the sender address is the most straightforward approach (Cranor and LaMacchia, 1998). However, the problem with this technique is that it is easy for senders to obtain new e-mail addresses or use fake e-mail addresses. Besides, defining spammers and separate them from the normal users represents another problem. Disputes can occur if mail servers improperly classify normal users into spammers and refuse to deliver mails sent by these normal users.

2.4 Filtering by Address Validity

The format of e-mail is described in the Requests for Comments (RFC) 822 initially designed by Crocker (1982). According to RFC 822, each e-mail must include the headers of “From” and either “To”, “CC”, or “BCC”. The sender address is listed in the “From” header and the receiver addresses are listed in the “To”, “CC”, or “BCC” headers. All e-mails should have at least a sender and a receiver.

Although RFC 822 mandates the existence of the "From" header in e-mails, no mechanism exists to ensure the validity of the sender address. That is, the e-mail address left in the "From" header can be invalid if the sender so chooses.

The validity of e-mail addresses of the header session may provide a cue for anti-spam filtering. The return addresses of unsolicited e-mail can be invalid for numerous reasons including that the spammers hope to avoid being sued if the e-mail breaks the law, disrupts Internet service or initiates an electronic "bombing" attack. Most e-mail servers suspend the email services of spammers. Moreover, some users who hate receiving unsolicited e-mails use electronic "bombing" to shut down the e-mail accounts of spammers.

The "To", "CC", and "BCC" headers are used to identify message recipients. "To" is used to identity main message recipients. Moreover, "CC" standing for Carbon Copy, identifies second message recipients rather than main message recipients. The function of the "To" and "CC" headers thus are very similar. However, "BCC" differs from the "To" and "CC" header. "BCC" means Blind Carbon Copy and blinds individual recipients to identify other recipients. No receivers thus can see the e-mail addresses of other "BCC" receivers. E-mail servers generally add "undisclosed-recipients" to the "To" header if only "BCC" is available.

Junk e-mails are sent in bulk and have numerous recipients. Spammers generally use the "BCC" header rather than "To" or "CC" for the receiver address to avoid revealing that the mail is part of a large batch. This habit of spammers provides a further cue for judging the possibility that an inbound e-mail is a junk e-mail.

3. EMPIRICAL STUDY

Content analysis can be used to examine sender and receiver addresses of junk e-mail. This study attempts to identify rules to discriminate desired and junk e-mails. Statistic analysis of header fields is performed for this purpose. The validity of the sender and receiver addresses also is checked to determine if they could be cues to filter out unsolicited e-mail. The checks used in this study include Domain Name Server (DNS) checking for the existence of the mail servers indicated in the e-mail address and Simple Mail Transfer Protocol (SMTP) checking for the existence of the e-mail accounts. E-mail messages were sent to check the existence of the accounts when the SMTP servers refused to respond to e-mail addresses validity checks.

3.1 Collecting Junk Mail

To gather unsolicited e-mails, an e-mail account was obtained from a Taiwanese Internet Service Provider (ISP), and bulk mail filtering options were not activated. This e-mail account was not used for any other purposes. All received mails except for messages sent by the ISPs system administrators were unsolicited. Unsolicited e-mails were collected over a twenty-month period between November 2001 and June 2003, yielding 3,417 junk mails.

3.2 Analysis of Sender and Receiver Addresses

Return address validity was checked for all junk mails. Figure 1 illustrates the results. Of the 3,417 junk mails collected, 277 (8.1%) had no sender address. The addresses of the remaining 3,142 e-mails were checked. Of this group, 178 (5.2%) were invalid because the domain name of the account mail server was unregistered, and a further 797 (23.3%) were confirmed to be invalid by SMTP checking. However, many SMTP servers refuse to respond to e-mail address checks to prevent spammers from collecting-mail addresses. Consequently, a real e-mail was sent to check whether the account existed. 806 (23.5%) e-mail addresses were confirmed invalid by obtaining

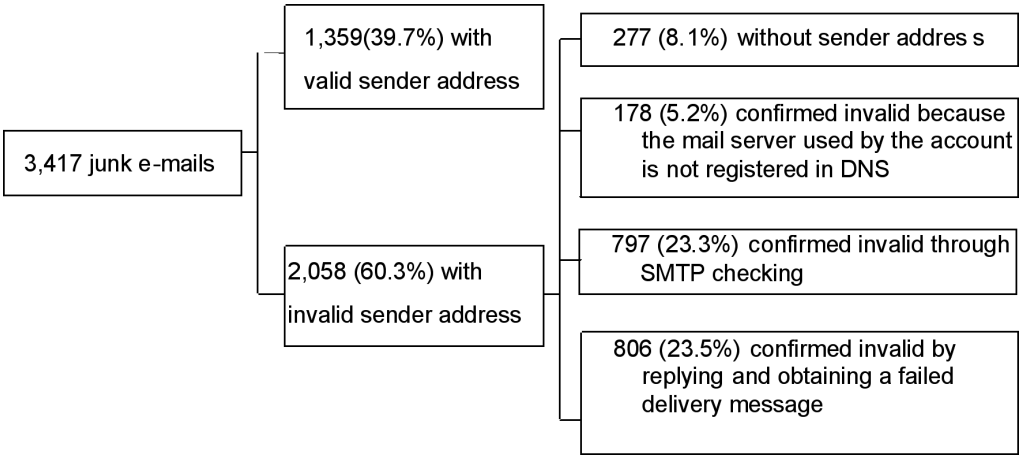

Figure 1: Sender Addresses of Received Junk Mail

warnings that the message could not be delivered. Overall, 2,058 (60.3%) of the 3,417 junk mails lacked a valid sender e-mail address.

Of the 3,417 junk mails, 247 displayed the receiver address in the "To" or "CC" header, while the remaining 3,170 (92.8%) displayed the receiver addresses in the "BCC" header. This strong tendency to use the BCC header may also provide a cue for judging whether an e-mail is junk. Although the fact that an e-mail places the receiver e-mail addresses in the "BCC" header does not mean that e-mail is definitely junk, such a placement may provide a cue for anti-spam filtering since almost all junk e-mail (92.8%) puts the receiver e-mail address in the "BCC" header.

4. DISCUSSION

The analytical results of this study confirm that many unsolicited e-mails have invalid sender addresses and put the e-mail addresses of the receiver in the "BCC" rather than the "CC" or "To" headers. The content analysis conducted here reveals that the validity of e-mail addresses left in the header session may provide cues for filtering spam. This study found that over 60.3% of unsolicited e-mails had an invalid sender address. Consequently, using the receiver address as a cue for spam filtering can filter out more half of all junk e-mails. This approach will not filter out ordinary e-mails out since typical e-mail users always include their true e-mail addresses to facilitate reply by receivers.

Additionally, this study also demonstrated that almost all junk e-mails place the e-mail addresses of the receivers in the "BCC" header. However, this habit of spammers is insufficient by itself to identify an e-mail as junk because the "BCC" header is used frequently by ordinary e-mail users, specifically to prevent the receivers from knowing the identities of other to whom the e-mail has been sent. However, this habit of spammers still may provide an additional cue for use in anti-spam filtering. This cue can be especially helpful in preventing legitimate mail from being classified as junk. If the receiver e-mail address appears in the "To" or "CC" headers, then the e-mail is unlikely to be junk is low, since almost all junk e-mail put the receivers' e-mail addresses at "BCC" header.

This study indicated that e-mail addresses in the header section could provide a cue for filtering junk e-mails. However, e-mail addresses in the header section are merely a supplementary or extra cue rather than a replacement for other filter out techniques. Applying the idea proposed by this

study in coexistence with other anti-spam filters would cause no conflict, and could boost the efficiency of anti-spam filtering.

REFERENCES

COHEN, W. (1996): Learning rules that classify e-mail, *AAAI Spring Symposium on Machine Learning in Information Access*, Palo Alto, USA: 18–25.

CRANOR, L. F. and LAMACCHIA, B. A. (1998): Spam. *Communications of the ACM* 41 (8): 74–83.

CROCKER, D. H. (1982): Standard for the format of ARPA internet text messages. *The Requests for Comments (RFC)* 822, http://www.rfc-editor.org. Accessed 15 July 2003.

NUCLEUS RESEARCH (2003): Spam: The silent ROI killer, *Research Note D59*, Nucleus Research. http://www.nucleusresearch.com. Accessed 01 July 2003.

PAZZANI, M. J. (2000): Representation of electronic mail filtering profiles: A user study, *Proc. the 5th International Conference on Intelligent User Interfaces*, New Orleans, LA, USA, 202–206, ACM Press.

POSTEL, J. (1975): On the junk mail problem. *The Requests for Comments (RFC)* 706. http://www.rfc-editor.org. Accessed 15 July 2003.

ROBINSON, G. (2003): A statistical approach to the spam problem. *Linux Journal* (107): 3.

SAKKIS, G., ANDROUTSOPOULOS, I., PALIOURAS, G., KARKALETSIS, V., SPYROPOULOS, C.D. and STAMATOPOULOS, P. (2003): A memory-based approach to anti-spam filtering for mailing lists. *Information Retrieval*, 6(1): 48–73.

SWARTZ, N. (2003): Spam costs businesses $13 billion annually, *Information Management Journal*, 37 (2): 9.

BIOGRAPHICAL NOTES

Chih-Chien Wang, PhD, is currently assistant professor of information management at the National Taipei University. His research interests include electronic commerce, online consumer behaviour, Internet rumours, e-mail spamming, and the impact of Internet to society. He also serves as the executive editor of Electronic Commerce Studies, a quarterly academic journal in Taiwan, and trustee council member of Association of Taiwan Electronic Commerce.

Chih-Chien Wang