

Toward Automated E-mail Filtering – An Investigation of Commercial and Academic Approaches

Jenq-Haur Wang and Lee-Feng Chien

Institute of Information Science, Academia Sinica,
Taipei, Taiwan, ROC

{jhwang, lfchien}@iis.sinica.edu.tw

摘要

本論文的目的在於探討如何發展出一套有效率又能有效自動過濾電子郵件的系統。為了更清楚地描述電子郵件過濾的機制，我們希望整合現有系統常用的 heuristic-based 與學術研究上常見的 classification-based (文件分類) 的方法，並且進一步發展出 Web-based 機器學習方法。

關鍵詞：電子郵件過濾，文件分類，垃圾郵件偵測，個人化

Abstract

The purpose of this paper is to investigate several major issues in developing an efficient and effective e-mail filtering system. To clearly describe the design of an e-mail filtering mechanism, we attempt to present an approach, with which some existing heuristics-based and classification-based techniques can be integrated, and more advanced techniques such as Web-based machine learning methods can be developed.

Keywords: E-mail Filtering, Text Classification, Spam Detection, Personalization

1. Introduction

E-mail has become one of the most popular communication tools while unsolicited bulk e-mails (also known as e-mail *spam*) are bringing more annoyance to most of the people using computers. Although anti-spam e-mail filtering systems and tools prevail, most users are still wasting much time and cost in computing resources and bandwidth dealing with spam. Conventionally, e-mail filtering is often treated as a typical application of text classification in academic research. As the techniques for filtering evolve, so do the tactics for e-mail spam to “masquerade” as legitimate ones. Therefore, simple analysis on e-mail header and content as a mixture of structured and unstructured text is not enough. Since spam is always changing, new e-mails have to be continuously collected if the latest features of e-mail spam are to be captured.

The purpose of this paper is to investigate several major issues in developing an efficient and effective e-mail filtering system. Academic researches on e-mail filtering focus on the application of conventional text classification methods, while existing e-mail filtering tools rely heavily on the construction of human readable and easy-to-maintain heuristic keyword-spotting rules. Successful integration of the two could result in better precision and recall of e-mail classification performance. In this paper, we attempt to present an approach, with which some existing heuristics-based and classification-based techniques can be integrated, and more advanced techniques such as Web-based machine learning methods can be developed. The presentation of the approach is two-fold: one is to describe the design of an e-mail filtering mechanism, and the other is to introduce the related research work that is being conducted in Academia Sinica.

2. A Brief Introduction to E-mail Transfer Mechanism

As networking and communication technologies evolve, people start to communicate with each other using tools like e-mails and instant messaging tools where pure texts, pictures, audio, video, and interactive multimedia communications among people are possible. With a simple and convenient mail client, messages can be delivered reliably and quickly to the recipients. In fact, e-mail transfer mechanisms are designed and mostly used when reliable delivery of messages is of major concerns. It's the fast, simple, and reliable delivery of messages that makes e-mails one of the most popular communication tools on the Internet. Many useful features are, therefore, being added into existing e-mail applications, for example, e-mail filing and organization, e-mail classification, and e-mail filtering.

To further investigate the problems in an e-mail transfer mechanism, the basic operations of the network protocol for e-mail delivery, Simple Mail Transfer Protocol (SMTP) [13], and the lifecycle of e-mails are illustrated in Fig. 1.

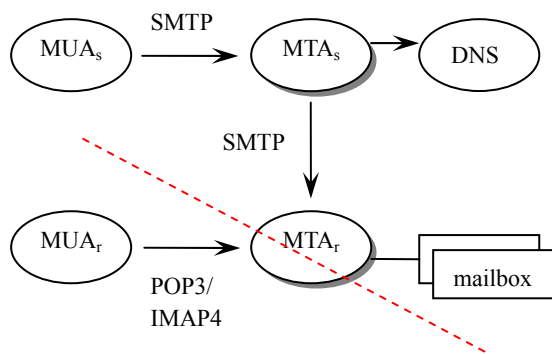


Fig. 1: An abstract diagram showing the lifecycle of e-mails in the process of a typical e-mail delivery mechanism.

In the process of e-mail delivery, there are two major components involved: MTA (Mail Transfer Agent) and MUA (Mail User Agent). As their names indicate, MTAs are mail servers or relays responsible for mail transfer between different domains, while MUAs are mail clients through which users can send and receive their e-mails. First of all, as shown in Fig. 1, the sender sends an e-mail with MUA_s which delivers the e-mail to MTA_s through SMTP protocol. After receiving the e-mail, MTA_s looks up for MTA_r in receiver domain from MX (mail exchanger) record in DNS (Domain Name System) [15] server, and forwards the e-mail to it. Then, on MTA_r, the e-mail will be stored in receiver's mailbox. When the receiver checks his/her e-mails, MUA_r will retrieve unread e-mails from MTA_r through POP3 (Post Office Protocol version 3) [17] or IMAP4 (Internet Message Access Protocol version 4) [6] protocol and finish the e-mail delivery process.

However, SMTP is so simple that many possible vulnerabilities could bring security problems to the users, such as forged e-mail headers, insecure e-mail transmission, and computer virus, among others. For example, an e-mail header can be forged such that the sender field contains an invalid e-mail address that the receiver could not distinguish its authenticity at all. Although security options such as SMTP Authentication [16] or SMTP over SSL/TLS [10] have been proposed to facilitate validation of the real identity of an e-mail sender by SMTP MTAs, not many MTAs implement this option, let alone correctly configure the options. Therefore, e-mail spam is becoming a serious problem that needs more attention.

With the growing use of e-mails, unsolicited commercial e-mails are also proliferating at an enormous speed. The most annoying part of e-mail spam is the *unsolicited* and often *bulk* nature, which is similar to the spread of computer virus. E-mail spam from companies advertising their products often arrives along with legitimate e-mails from your friends or colleagues. This could bring annoyance for individual users, industries, and ISP's. Users have to spend plenty of time manually examining each e-mail

and deleting uninteresting ones. Moreover, the productivity of employees will be greatly reduced, and the cost for ISP's will be increased since a large portion of network traffic is devoted to the delivery of e-mail spam. Therefore, e-mail filtering is necessary for protecting all users from being overwhelmed by growing volumes of junk e-mails.

3. Commercial Approaches to E-mail Filtering

There are many existing tools and systems for e-mail classification and filtering, for example, *SpamAssassin* [25], *popfile* [19], and *ifile* [21] are among the most popular ones. Although existing anti-spam tools prevail, the trend of increasing numbers of e-mail spam seems not to be slowed down at any rate.

Generally, there are two kinds of e-mail filters depending on where the filter is located: *client-side* e-mail filters on MUAs and *server-side* e-mail filters on MTAs. Since client-side e-mail filters have to be deployed on every mail client and unwanted e-mails will still get delivered from mail servers, server-side e-mail filters are more practical and effective in blocking junk e-mails. However, personalization support is better provided in client-side filters since user preferences can be more easily obtained. For instance, personal white list of intended correspondents can be easily stored as part of user profile on each mail client.

Usually, in real-world e-mail filtering tools, heuristic keyword-spotting rules are used in identifying possible e-mail spam. However, both filtering accuracy and automation level are often not satisfactory enough for users to widely deploy such filters in their mail clients. Other techniques include RBLs (Real-time Blackhole Lists) [20] where a black list of mail servers or relays for possible spammers are recorded in a DNS server. E-mails originating from those mail servers could very possibly be spam. Although RBLs could mark a whole domain as invalid, other valid users in the same domain would also be considered as spammers. White lists which explicitly mark intended correspondents or known e-mail addresses as valid senders are often adopted. Techniques for preventing open mail relays (MTAs that allow anyone to send e-mails through them) and open proxy servers (where every one can send any packets through them) have also been proposed so that only authorized users are permitted to utilize the services.

4. Classification-based Approaches to E-mail Filtering

In this paper, we are interested in investigating several major issues encountered in classification-based approaches to e-mail spam filtering. Generally, classification-based approaches to e-mail filtering can be divided into two kinds according to the number of categories for classification: *binary* classification and *general* classification. For binary classification, the filtering system only needs to decide if the incoming e-mail is spam or not. It's more practical to distinguish between e-mails that are *interesting* or *non-interesting* to a specific user since a person's spam may be useful information for another. On the other hand, general classification deals with problems such as filing e-mails into different folders according to the topics and user interests in the contents, where spam could be organized into one of many possible classes or categories. Since binary classification is a special case of general classification, we will focus on binary classification. Although multiple categorizations where each e-mail belongs to one or more categories could easily happen in real cases, in our discussion, single categorization for each e-mail will be assumed.

Existing researches on e-mail classification treat e-mails as a combination of structured text, such as the sender and subject fields, and unstructured text, such as the mail content. Conventional text classification methods such as Term Frequency-Inverse Document Frequency (TF-IDF), Naïve Bayes (NB), and Support Vector Machine (SVM), are then applied to the classification or filtering of e-mails.

However, since spam is always changing, we have to keep collecting new spam in order to extract dynamically shifting features from them. If we could not collect enough up-to-date e-mail spam as our corpus, it will be difficult to train the models of e-mail spam in terms of their ever-changing features.

Secondly, mail contents in spam can vary from "make-money-fast" schemes to weight loss programs or anti-aging products. Moreover, spam nowadays can masquerade as legitimate ones, just like an e-mail from your friend telling you about something new or interesting on a particular website. Therefore, mail contents are not always trustable for the purpose of classification. Classification through text analysis of only mail header and contents is not enough.

With the limited spam corpus and changing e-mail contents, conventional approaches cannot precisely capture the characteristics of e-mail spam. A further introduction to previous works is given below.

4.1. Text Classification and E-mail Filtering/Classification

Existing researches on e-mail filtering/classification focus on the application of conventional text classification methods. For example, Cohen [5] compared the accuracy of *Ripper*, a rule-based algorithm, and *Rocchio*, a TF-IDF method, in classifying mails. Sahami et. al. [22] proposed using Naïve Bayesian (NB) approach to filtering junk e-mails. Brutlag et. al. [2] compared SVM, TF-IDF, and unigram model in general e-mail classification. Diao et. al. [7] proposed comparing NB with decision trees (C4.5) [14] in classification-based personal e-mail filtering. Other approaches have also been proposed to combine existing methods in e-mail classification, for example, *boosting* algorithm [3].

4.2. E-mail Filtering System Techniques

Real-world e-mail filtering techniques for spam detection rely on the identification of the most common characteristics of e-mail spam. Heuristic keyword-spotting rules are usually used. Techniques utilizing distributed knowledge of e-mail spam are becoming more popular since spam usually arrives in bulk volume. It's similar to the Distributed Checksum Clearinghouse (DCC) [8]. Examples include Cloudmark's SpamNet [27], which is derived from Vipul's Razor [29], and a peer-to-peer spam filtering system called SpamWatch [28] in the OceanStore project [18]. Real-time Blackhole Lists (RBLs) [20] try to blacklist those spammers using only source e-mail addresses or domain names in e-mail headers. HoneyPot systems and projects such as HoneyNet [11] are also related to the idea of distributed e-mail collection and spam detection scheme.

5. Detailed Design of an E-mail Filtering Mechanism

For introducing the detailed design of an e-mail filtering mechanism, in this section we present an approach, which considered training the mechanism through learning of extra resources such as personalized e-mail collection and the Web.

5.1. Architecture

As shown in Fig. 2, the architecture for automated e-mail filtering and the interactions among MTAs and MUAs is illustrated.

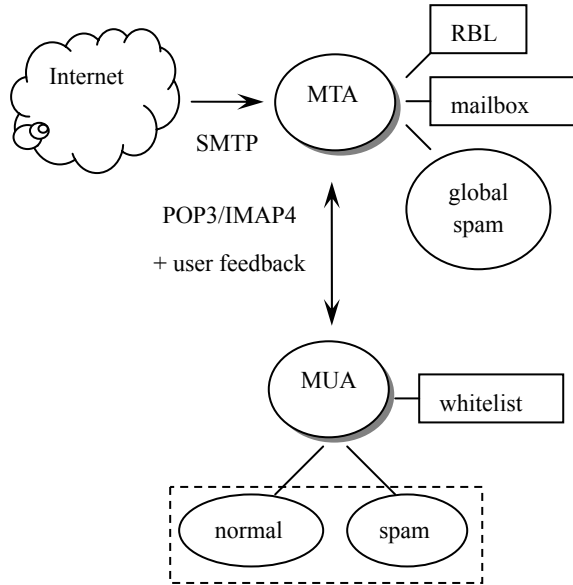


Fig. 2: An abstract diagram showing the proposed architecture for automated anti-spam e-mail filtering and the interactions between MTAs and MUAs.

On the MTA, global knowledge of e-mails, e.g., RBLs and global spam for the whole domain can be stored. On the other hand, local knowledge of e-mails such as a personalized white list and personal collection of both normal and spam e-mails can be stored on the MUA. The network protocol between MTAs and MUAs can be POP3 [17] or IMAP4 [6]. In addition, user feedbacks from MUA such as the personal markings of spam or normal e-mails have to be transmitted to MTA in order to provide better e-mail filtering support on server-side (MTA). Therefore, a modification to POP3/IMAP4 protocol or a proprietary protocol for user feedback is required in our proposed architecture.

5.2. Corpus Collection

A spam corpus can be built in MTA using a distributed e-mail collection scheme. Although public spam corpus such as Ling-Spam and PU1 corpus [1], SpamBase [26] from UCI Machine Learning Repository, and SpamArchive [24] are also available, we can still use the distributed collection scheme, which is similar to p2p spam filtering. It can be considered to create another separate corpus for both spam and normal e-mails that is built from user's personal mailbox.

Since it's possible to receive e-mails in both English and other languages like Chinese, different features could be learned from spam in different languages. Multilingual e-mails are possible in non-English speaking countries, where English terms usually appear as a named entity. Therefore, general

concepts that are independent of languages can possibly be learned from a bilingual dictionary. However, dictionary is usually not enough since many marketing terms often utilize the latest terminologies from the news. For example, during the SARS (Severe Acute Respiratory Syndrome) outbreak in Taiwan, commercial e-mails began advertising healthcare products related to the prevention of SARS infections. Therefore, extra resources such as personal e-mail collection and the Web could be quite helpful. Personal e-mail collection usually contains both positive and negative instances of e-mails that could be useful in identifying personal interests of various topics in e-mails. On the other hand, lots of information on the Web is up-to-date and multilingual, especially in the so-called Chinese Web, which could provide valuable information if appropriate Web mining is done and useful patterns extracted.

5.3. Pattern Extraction

Once the e-mail collection is built, the second task is to extract patterns and learn concepts of spam from the corpus. Since the mail contents could be diverse, instead of directly calculating the similarity between e-mails, we want to identify text patterns that are representative in e-mail spam. Therefore, pattern extraction methods such as *PAT-tree based* approach [4] can be used. Firstly, text patterns are extracted from e-mails in the training corpus for global spam. Then, the extracted text patterns are used to represent the features of these patterns. In our approach, a conventional vector space model for text representation is applied to these text patterns. For example, a TF-IDF weighting scheme [23] as follows can be used.

$$w_{ij} = tf_{ij} * \log \frac{N}{n} \quad (1)$$

where w_{ij} is the weight of term T_j in pattern P_i ,
 tf_{ij} is the term frequency of term T_j in pattern P_i ,
 N is the number of patterns in the collection,
 n is the number of patterns where term T_j occurs.

Therefore, a feature vector can be derived for each e-mail and used as the "characteristic" of this e-mail.

5.4. Feature Selection

Note that the number of features could be very large in the text patterns extracted from e-mails. In order to reduce the dimension of feature space and also the computational complexity, feature selection methods [30] should be applied. For example, Mutual Information (MI) and Information Gain (IG) are two of the most common methods for feature selection.

5.5. Distance Measure

In order to effectively distinguish between e-mails with various degrees of ‘spamminess’, a similarity measure between two feature vectors can be defined as follows:

$$\text{sim}(v_a, v_b) = \cos(v_a, v_b) \quad (2)$$

Basically, we can use the *cosine* measure where the angle between two vectors in the feature space is measured.

5.6. Clustering Algorithm

Finally, a hierarchical clustering algorithm like HAC (Hierarchical Agglomerative Clustering) [12] can be applied to these feature vectors and a hierarchy of clusters corresponding to different concepts of e-mails can be obtained. In our approach, two separate topic hierarchies for legitimate e-mails and spam will be obtained.

5.7. Classification of New E-mails

When a new e-mail arrives, we can do the same as in the process of training the concepts of spam. First of all, a feature vector is derived from each e-mail using a TF-TDF weighted vector of extracted text patterns. Then, a classification algorithm like *kNN* (k-nearest neighbor) [14] should be used for measuring the nearest concept of each e-mail. Since concepts of e-mail spam are quite personal, we could classify an e-mail into the two topic hierarchies for global spam and local ones, respectively.

6. Discussions

There are several issues when dealing with e-mail filtering/classification. First of all, the reduction of *false-positive* rate is as critical as the accuracy (precision, recall) of classification. Note the cost-sensitive characteristic in e-mail classification: *false-positives* where legitimate mails are caught as spam are considered more unacceptable than *false-negatives* if a few spam are not caught. Secondly, the runtime efficiency of classification can be critical in real-world systems since real time filtering is required. Therefore, incremental adaptation of the algorithm could help improve performance. Finally, the size of training set, feature set size, and class sparseness could also affect the accuracy of classification which need further investigation.

According to the direction of e-mail transfers, the purpose of filtering could be different. For example, filtering on *incoming* e-mails focuses on dealing with reducing the number of e-mail spam while filtering on *outgoing* e-mails concerns more about auditing e-mails that may leak confidential information in a

company without permission. Since our approach focuses on extracting text patterns and learning concepts of topics in an e-mail, both directions of e-mail transfers can be handled as long as the definition of spam and selected features can be adjusted as needed.

Table 1 shows the possible types of e-mails according to the relationships between sender and receiver.

Receiver/Reply Sender Type	Never Receive	Received, but No Reply	Received and Replied
Sender in Whitelist	Not (First Contact)	Not	Not
Sender in Blacklist	Spam (Known spammer)	Spam	Not (exception)
Unclassified Sender	Unknown	Not Interested	Not

In our analysis, e-mails can be divided into several types according to the sender/receiver relationship as shown in Table 1. Factors that affect the categories include whether the sender e-mail address is contained in blacklists and white lists, whether e-mails have been received from the sender, and whether previous e-mails from that sender have been replied. First of all, if the sender is in the white list, his e-mails will be legitimate no matter they have been replied or not. Secondly, if the sender is in the blacklist, his e-mails will be considered as spam unless previous e-mails have been replied. Replying the e-mail from a sender in the blacklist may indicate an exception to the blacklist. Finally, if the sender is neither in the blacklist nor the white list (the unclassified sender), we have to check if we have received mails from him or not. If this is the first time he send you an e-mail, the status of his e-mail will be unknown since no prior information can be obtained for that sender. If we did receive and reply his previous mails, then the e-mail should be legitimate. If we didn’t reply his previous e-mails, maybe we are not interested in the topics he wrote. All possible combinations are listed in Table 1.

Therefore, user feedbacks from MUAs could be important indications on determining whether an e-mail is legitimate or spam according to the types of senders and the receiver/reply relationships.

7. Conclusion

The purpose of this paper is to investigate several major issues in developing an efficient and effective e-mail filtering system. Successful integration of the

commercial approaches and academic approaches could result in better precision and recall of e-mail classification performance. We have attempted to present an approach, with which some existing heuristics-based and classification-based techniques can be integrated, and more advanced techniques such as Web-based machine learning methods can be developed.

References

- [1] I. Androutsopoulos, et. al, "An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages," in *Proc. SIGIR 2000*, pp. 160-167, Jul. 2000.
- [2] J. D. Brutlag and C. Meek, "Challenge of the E-mail Domain for Text Classification," in *Proc. ICML 2000*, pp. 103-110, 2000.
- [3] X. Carreras and L. Marquez, "Boosting Trees for Anti-Spam Email Filtering," in *Proc. Euro Conference on Recent Advances in Natural Language Processing (RANLP 2001)*, Sep. 2001.
- [4] Lee-Feng Chien, T. I. Huang, and M. C. Chien, "PAT-tree-based Keyword Extraction for Chinese Information Retrieval," in *Proc. SIGIR 1997*, pp. 50-58, Jul. 1997.
- [5] W. W. Cohen, "Learning Rules that Classify E-Mail," in *Proc. AAAI 1996*, pp.124-143, Mar. 1996.
- [6] M. Crispin, "Internet Message Access Protocol – version 4rev1," *RFC 3501, IETF*, Mar. 2003.
- [7] Y. Diao, H. Lu, and D. Wu, "A Comparative Study of Classification Based Personal E-mail Filtering," in *Proc. PAKDD 2000*, pp. 408-419, Apr. 2000.
- [8] Distributed Checksum Clearinghouse (DCC), available at: <http://www.rhyolite.com/anti-spam/dcc/>
- [9] H. Drucker, D. Wu, and V. N. Vapnik, "Support Vector Machines for Spam Categorization," *IEEE Transactions on Neural Networks, Vol. 20, No. 5*, Sep. 1999.
- [10] P. Hoffman, "SMTP Service Extension for Secure SMTP over Transport Layer Security," *RFC 3207, IETF*, Feb. 2002.
- [11] The HoneyNet Project, available at: <http://project.honeynet.org>
- [12] A. Jain and R. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.
- [13] J. Klensin, ed., "Simple Mail Transfer Protocol," *RFC 2821, IETF*, Apr. 2001.
- [14] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
- [15] P. Mockapertris, "Domain Names – Implementation and Specification," *RFC 1035, IETF*, Nov. 1987.
- [16] J. Myers, "SMTP Service Extension for Authentication," *RFC 2554, IETF*, Mar. 1999.
- [17] J. Myers and M. Rose, "Post Office Protocol – Version 3," *RFC 1939, IETF*, May 1996.
- [18] OceanStore project, available at: <http://oceanstore.cs.berkeley.edu/>
- [19] popfile, available at: <http://popfile.sourceforge.net/>
- [20] Real-time Blackhole List (RBLs), available at: <http://mail-abuse.org/rbl/>
- [21] J. D. M. Rennie, "ifile: An Application of Machine Learning to E-mail Filtering," in *Proc. KDD 2000*, Workshop on Text Mining. Aug. 2000.
- [22] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail," in *Proc. AAAI 1998*, Jul. 1998.
- [23] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management, Vol. 24, No. 5*, pp. 513-523, 1988.
- [24] Spam Archive Project, available at: <http://spamarchive.org/>
- [25] SpamAssassin, available at: <http://spamassassin.org/>
- [26] SpamBase, Spam e-mail database from UCI Machine Learning Repository, available at: <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [27] SpamNet, available at: <http://www.cloudmark.com/products/spamnet/>
- [28] SpamWatch, available at: <http://www.cs.berkeley.edu/~zf/spamwatch/>
- [29] Vipul's Razor, available at: <http://razor.sourceforge.net/>
- [30] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in *Proc. ICML 1997*, pp. 412-420, Jul. 1997.