IMPROVING ANTISPAM TECHNIQUES BY EMBRACING PATTERN-BASED FILTERING

HAIRUL ANUAR BIN MAT NOR

UNIVERSITI TEKNOLOGI MALAYSIA

IMPROVING ANTISPAM TECHNIQUES BY EMBRACING PATTERN-BASED FILTERING

HAIRUL ANUAR BIN MAT NOR

A project report submitted in partial fulfilment of the

requirements for the award of the degree of

Master of Science (Information Security)

Centre for Advanced Software Engineering

Faculty of Computer Science and Information System

Universiti Teknologi Malaysia

APRIL 2009

ABSTRACT

This study attempted to show that there are still away to improve antispam system. The classical method of filtering spam is by inspecting content of an e-mail and finding a matching pattern with a predefined ruleset. Each matched keyword or sentence will produce a weight, also called score, which will be combined to produce the final score and later to be used for identifying spam similarities in the message. Spammers keep changing a style in generating a spam message to avoid being filtered. Bayesian technique was found to be suitable to embed in the antispam in order to recognize the characteristic of spam in a message eventhough the content has been changed. However, the Bayesian introduced difficulties to the system as spammers have changed the way they send the spam specifically to bypass Bayesian filter. Thus, it is time to find a way to filter those spam. Normally an antispam works on the content, but there is a possibility to filter spam based on its pattern of delivering the spam at network level to reduce the congestion in the network. Filter at network level is also benefiting the server as it has eliminated some spam before they are received and processed for the content. A study were conducted to show above statement is true. An Antispam with Pattern Based Filter (ASPBF) will be tested and the result will be compared with the test for Antispam with Bayesian. The comparison result will determine how much it has achieved its objectives. This study will be able to stimulate more studies in the future to further improve antispam solution in the fight against spam and to have a better e-mail communications.

ABSTRAK

Kajian ini dilakukan untuk membuktikan bahawa masih terdapat peluang untuk memperbaiki sistem *antispam*. Cara biasa mengatasi masalah spam adalah dengan memeriksa kandungan emel dan mencari kesamaannya dengan himpunan maklumat kandungan *spam* yang tersedia ada. Ini telah memaksa para pembuat *spam* mengubah cara mereka mengeluarkan *spam* bagi memastikan ia tidak ditapis. Susulan daripada itu, para penyelidik mendapati teknik yang digunakan Bayesian adalah sesuai untuk menjejaki emel yang mengandungi ciri-ciri *spam* walaupun isi kandungan e-mel tersebut diubah bagi mengaburi *antispam*. Tetapi, teknik Bayesian didapati tidak lagi releven malah telah mengakibatkan keburukan yang lain terutama sekali bila para pembuat *spam* mengubah kandungan emel semata-mata untuk mengaburi teknik Bayesian. Sungguhpun begitu, *spam* masih boleh dikenal pasti melalui cara ia dihantar, oleh itu mengecam corak *spam* dihantar pada peringkat rangkaian mampu mengekang kepadatan rangkaian dan mampu mempertingkatkan proses *antispam* kerana kebanyakan e-mel telah disekat dan tinggal sebilangan kecil sahaja yang akan diproses oleh antispam yang mengkaji kandungan e-mail. Kajian ini juga membangunkan satu aplikasi *antispam*, dinamakan *Antispam with Pattern-Based Filter* (ASPBF), berupaya mengecam corak penghantaran spam dan memutuskan sambungan spam tersebut. Hasil ujian ini akan membuktikan ia telah berjaya memenuhi objektifnya dalam mempertingkatkan kelancaran dan keberkesanan *antispam*. Hasil ujian ke atas aplikasi tersebut dan aplikasi *antispam* dengan Bayesian akan dibandingkan bagi menunjukkan keberkesanan aplikasi tersebut. Kajian ini diharapkan bakal menjadi perangsang kepada kajian-kajian lain dalam mempertingkatkan kemampuan *antispam* bagi memperoleh dunia komunikasi e-mel yang lebih baik.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

This chapter forms the introduction to the thesis. Discussions begin with emphasis on the history of spam and ways taken to fight against spam, including the use of Bayesian technique, which is later found to be causing another problem.

1.2 Evolution of Spam and Ways to Filter

The growth of technologies especially those related to internet usage has gradually been producing side effect that are damaging one's productivity and image when a precaution steps are not taken seriously. One of it is spam, less internet users aware the severity of damage from spam. In addition, there are many users do not know anything about spam, yet faced it in daily business.

Zdziarski (2005) claimed that many people described spam as unwanted e-mails, while others defined it as "unsolicited commercial e-mail" which covered advertisements for products and other types of solicitations. Eventually, the meaning was leaning towards "unsolicited bulk e-mail" and Zdziarski (2005) suggested it was more acceptable.

It is commonly known that spam is sent in bulk intended to reach as many people as it could to deliver messages that are meant to solicit. Therefore, the beginning of mass spam distribution could be traced back to year 1994.

Lawyers Laurence Canter and Martha Siegel implemented the first large-scale spam in 1994, especially when they sent a message offering help with the Immigration and Naturalization Service's "green-card lottery" to thousands of Usenet newsgroups. Thousands of angry recipients sent nasty messages to NETCOM, Canter and Siegel's Internet Service Provider, causing the provider's machine to overload and crash (Foxman and Schiano, 2000).

Since then, people started to protect their e-mail servers from spam, thus creating a huge demand on antispam products. This had motivated companies to produce antispam solutions in rapid mode by taking advantage from readily available antispam software developed by open source community.

One of the popular open source antispams is called SpamAssassin which was developed by using Perl language and the implementation was meant for UNIX or Unix-like environment.

Schwartz (2004) described SpamAssassin as a software for analyzing e-mail messages, determining how likely they were to be spam and reporting its conclusions. It was a rule-based system that compared different parts of e-mail messages with a large set of rules. Each rule would add or remove points from the message's spam score. A message with a high enough score would be reported as a spam.

There are numbers of rules embedded within SpamAssassin which contribute scores to the final decision. Each rule has its own weight where the most effective rules will be given more weight. However, users are also given chances to modify the score of each rule or to add new rule sets.

SpamAssassin combines message format validation, content-filtering and the ability to consult network-based blacklists (Schwartz, 2004). These are the typical ways to filter spam. In general as described by Mendez (2008), there are two kinds of spam filtering techniques: (i) collaborative system and (ii) content-based approaches. The former are based on sharing identifying information about spam messages within a filtering community, while content-based approaches analyzed message content in order to identify its class (usually using machine learning techniques)

In the early age, spam could be found in a typical way. However, as time goes by spammers were getting more creative and they produced spam in different ways making the spam more dynamic and less chances to be filtered by antispam. Thus, artificial intelligence was needed in order to countermeasure the dynamic spam attacks. That was why Bayesian techniques were embedded with SpamAssassin to improve the catch rate.

Bayesian Filter in SpamAssassin is one of the most effective techniques for filtering spam. It is a mathematical statistical analysis. Bayesian analysis involves teaching a system that a particular input gives a particular result. For spam filtering, this teaching is repeated, many times over with many spam and ham (legitimate) e-mails. Once this is finished, a Bayesian system can be presented with a new e-mail and will give a probability of the result being spam. However, for best results, teaching should be a constant process (McDonald, 2004).

1.3 Problem Statement

As enhancements are continuously introduced to antispam solutions and Bayesian was always the main focus for its capabilities to detect spam using probabilistic approach. However, amount of traffic and amount of spam increased over time causing Bayesian technique to undergo performance degradation. In relation to Bayesian technique, this section will primarily discover the limitations of Bayesian technique.

1.3.1 Limitation of Bayesian

Bayesian technique used in antispam is primarily meant for probability and similarity of sentences and keywords with actual spam. Therefore, Bayesian can only be used for spam that is readable by a machine, which means words are in correct spelling sequence.

A problem appeared when spammers started to use non-readable words, where other characters are introduced in the middle of the spelling. Machines will face problem recognising the words, but human eye will be able to identify the words as the words and the unknown characters are differentiated by colours or by any other mean. This is where Bayesian is gradually failing its operations.

More unfortunate to Bayesian filter, spam attacks not only can slip through the filter due to the unreadable form of message, latest research has shown that readable form of message also can defeat the Bayesian filter.

Research by Karlberger et al. (2007) had shown that existing attack aiming at Bayesian is to let spam mails be identified as ham mails. Currently existing attacks aim to achieve this by adding words to the spam mails. The objective is that these additional words are used in the classification of the mail in the same way as the original words, thereby tampering with the classification process and reducing the overall spam probability. When the additional words are randomly chosen from a large set of words, this is called random word attack, also called “word salad”. The objective is that the spam probabilities of words added to the spam message should compensate for the original tokens’ high spam probabilities in the calculation of the whole message’s combined spam probability.

This has shown that spammers has turned significant energy to building software which cloaks their spam to hide the trigger words and phrases from Bayesian filters, the resulting messages become progressively simpler for pattern based scanners to correctly identify as spam. This is because pattern based scanners can identify spam by actually targeting the cloaking techniques the spammers are using, rather than worrying about identifying spam based on words and phrases.

There are significant patterns that can be recognised from a spam e-mail as spam is normally sent in bulk targeting a group of people. Therefore, there are many similar spams are flowing around in the internet. Thus, it is a good idea to take advantage on this characteristic to enhance the filtering process.

However, not just pattern in the spam e-mails are always similar, but techniques of sending the spam also found to be similar. This has increased

awareness among antispam developer to look at filtering the spam before the spam actually reach the e-mail server, thus saving a lot of processing power.

These two types of pattern are not the centre of Bayesian's focus even though it can be of assistance to an antispam to perform better. In the event some spams will slip through the pattern-based filter, Bayesian can be used as the last line of defence.

1.3.1.1 Image Spam

Image spam is another form of spam due to its effective ways to deceive spam filters since they can only process text (Mehta et al. 2008). Images are hard to process for identifying spam as they only contain non-readable characters, which can only be interpreted by human's eye. Figure 1.1 below show one example of image spam.

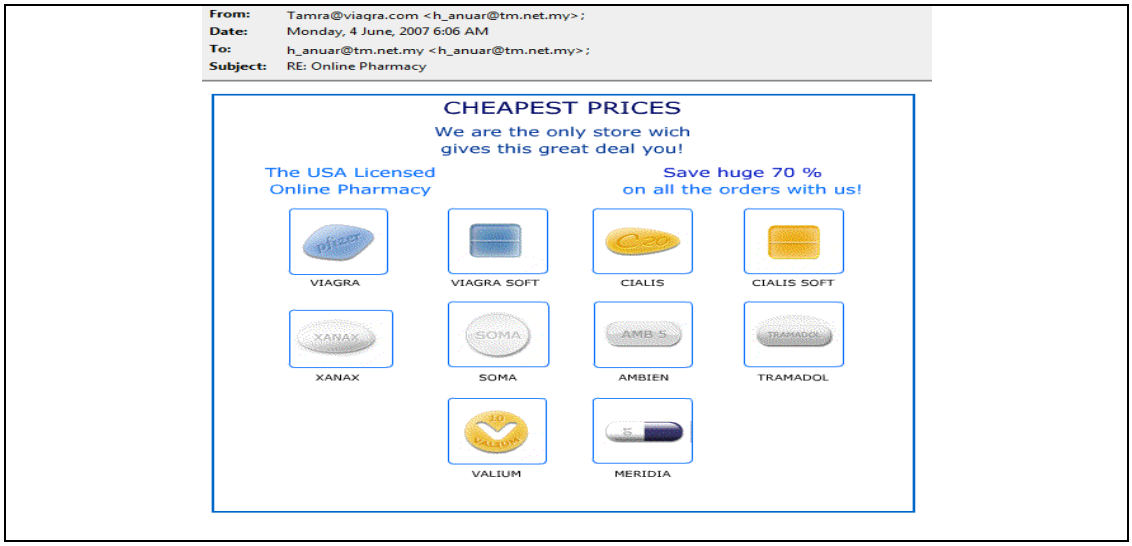

Figure 1.1: Sample of Image Spam

Although Optical Character Recognition (OCR) was introduced to extract letters and keywords from an image, but it has given a big impact on overall processing load.

Mehta et al. (2008) also emphasis that images in spam e-mails contain text messages conveying the intent of the spammer. This text is usually an advertisement and often contains text, which has been blacklisted by spam filter.

Since legitimate e-mails may also contain images in the attachment, therefore it is hard to differentiate legitimate e-mail and spam e-mail if the detection cannot filter spam in an image. Besides, detecting characteristic of the e-mail with image spam may not the best solution as the characteristics are almost similar to the legitimate e-mails.

In further research by Baskhar et al. (2008), they found the sophisticated techniques like Optical Character Recognition (OCR) can fail to recognize text in images due to the distinction of every image spam, where more noises where added, random background, colour of the fonts and many more.

1.3.1.2 Open relay and proxies

Open relay is a method to provide a convenience way to end users enabling them to send e-mails without any authentication that require manual intervention each time. This method also simplifies communication between internal server and the upstream server that connect to the internet.

However, open relays have been exploited by spammers due to the anonymity and amplification offered by the extra level of indirection. It appears that the widespread deployment and use of blacklisting techniques have all but extinguished the use of open relays and proxies to send spam (Ramachandran and Feamster, 2006).

Although smtp authentication is widely used nowadays to prevent unauthorized use of the e-mail server, but there are still a lot of e-mail providers that are not taking serious action to overcome this issue as it does not really affect their servers.

That is not the only problem, but spammers are developing their smtp server, a tiny version, in order to find a simpler way to broadcast the e-mail rather than finding an open relay servers which may take time to discover. This tiny smtp server is then embedded into botnet, therefore they will have a large distribution of open relay servers that are ready for use at anytime.

1.3.1.3 Botnets

Today many people claimed that most of the spams are coming from botnet, where botnet is an interpretation of a collection of machines acting under one centralized controller. Normally a machine might get affected by worms that exploiting Windows vulnerabilities opening backdoor to allow remote attacker to take control of the infected machines.

As mentioned earlier, the botnets might have a tiny smtp server application, thus whenever the backdoor is opened to the attackers, they will have a direct access to the smtp server to launch spam attacking everyone in the world. This will hide the real spammers, but put the blame on the innocence.

But not necessarily the botnets to have its own smtp server, it can exploit vulnerabilities in the windows machine to act as smtp server. As mentioned by Ramachandran and Feamster (2006), worms exploiting DCOM and LSASS vulnerabilities on windows systems allows infected hosts to be used as a mail relay, and attempts to spread itself to other machines affected by aforementioned vulnerabilities, as well as over e-mail.

1.3.1.4 Spam With Dynamic Contents

Content-based filters, such as those incorporated by popular spam filter like SpamAssassin, successfully reduce the amount of spam that actually reaches a user's inbox. On the other hand, content-based filtering has drawbacks. Users and system administrators must continually update their filtering rules and use large corpuses of spam for training; in response, spammers is negligible, since spammers can easily alter content to attempt to evade these filters (Ramachandran and Feamster, 2006).

Although Bayesian technique was initially found to show high performance precision and recall, it has several problems. First, it has a cold start problem, that is training phase has to be done before execution of the system, the system has to be

trained about spam and non-spam mail. Second, the cost of spam mail filtering is higher than rule-based system. Third, if an e-mail has only few terms those represent its contents, the filtering performance is fallen (Kim et al, 2004).

In that sense, the Bayesian works based on similarities to the previously found spam and non-spam. However, the newly designed spam may not be recognized by Bayesian, thus require a human intervention to provide an input for re-training and to get familiar with the latest content. But, while waiting for a human input, damages might have been done and reputation of the e-mail providers might get poorer.

1.3.2 Available Solution in Resolving Issues with Bayesian

This study was inspired by a solution commercially made available, namely Extensible Messaging Platform (EMP), which is focusing on recognising spam from its pattern and signature. Although the solution is ready off-the-shelf, but there are some characteristics that may affect performance and reliability.

Figure 1.2 below shows the architecture of the EMP solutions:

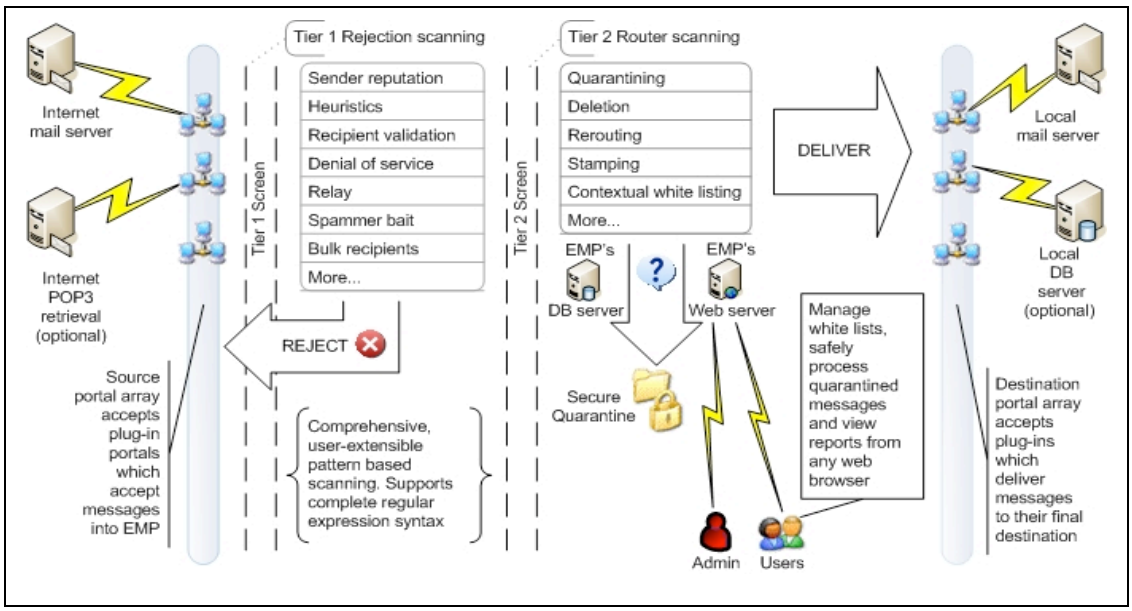

Figure 1.2: Architecture of EMP

From the given architecture, there are some factors that may lead to a decreasing performance:

i. The technique used to do sender reputation is based on blacklist database which information are collected and shared from various input all over the world, thus requesting the information in real-time requires a reliable connections. This is a very resource intensive process especially when the amount of the traffic is too huge.

ii. The EMP is acting as an smtp software which create more load to the server as it needs to run two instances of different smtp software.

iii. The signature of spam shall be downloaded from the central repository in order to make the pattern-based filter to function properly, it would be better if the machine is self-dependent where it shall develop its own database of signature from the observation and analysis on the incoming traffic.

Although a commercial product is available to provide pattern-based filter, this study will look deeper into ensuring the product to maintain its performance for a longer period of time even though the amount of traffic increases.

Besides, this is an academic study to appreciate researches by earlier researchers for their contribution in improving antispam solution. This study shall be more valuable and acceptable when each of the proposed technique in the past research is put into a test.

1.4 Definition of Pattern

The use of ‘pattern’ is referring to the following definitions which depend on which protection level it may be used:

i. At network level

At this level, the pattern referring to behaviour of an IP address in requesting for SMTP connections. Some of the patterns are:

- The large amount of traffic in a short period of time.
- Number of invalid recipient from an IP address in a short time.
- Invalid IP address or IP address without specific DNS record.

ii. At e-mail content level

At this level, the some characteristic, especially those listed below will be categorized as pattern:

- Invalid timestamp.
- Invalid header information.
- Keywords or mix of keywords matched with database of spam keywords.

iii. At e-mail attachment level

The following items are the pattern to look for:

- Image containing spam keywords. The pattern can be figured out by recognising the properties of the image and compare it with template of image spam.

1.5 Objective

This study will be focusing on three main objectives:

i. To study other solutions to improve the percentage of catch rate and to reduce processing load.

This study will try to look for other antispam solutions which are avoiding Bayesian in its implementation. Bayesian is known with its probabilistic approach which requires a lot of sample data and yet intensive calculation for a higher accuracy. With a large data and calculation it will require more processing power which reduces overall performance of an antispam product. Thus, a simple way known as pattern-based filtering will serve a better performance as its functionality does not need to thoroughly look into the content of an e-mail.

ii. To analyse combinations of existing solutions to form the best pattern-based filtering technique.

There are studies done by researchers that tried to improve anti spam filtering without using Bayesian filter technique in the implementations. However, in order to prove they had improved the solutions, they had to compare their results with the result from antispam product with Bayesian filter. However, each of these techniques was only focusing on specific function and parameter. Therefore, this study will help to determine the right combination of techniques that could potentially improve the antispam solution without the need to include Bayesian filter.

iii. One antispam system will be developed at the end of this study. The said system will be formed by a combination of selected techniques.

The said system will be developed without Bayesian. But, for a comparison purposes, result from an existing product that is using Bayesian will be used to compare with the result of the newly developed system to prove that an improvement has been made.

1.6 Scope

This study is introducing techniques to improve antispam, especially to reduce the erroneous and the cost of performance generated by Bayesian filter. Although there are numbers of techniques focusing on other areas of improvement, this paper will not study the foundation of those techniques.

This study will further study on the suitability and compatibility of each of mentioned techniques in order to form the pattern-based filtering technique to be used in an antispam solution or product. Having the right tool and a controlled environment, the said antispam system will be developed by using pattern-based filtering, and will be tested for a comparison study against other antispam solution that is using Bayesian.

The proposed antispam system shall have three levels of protection where the pattern-based filtering can be applied:

i. Network Level
ii. E-mail Content Level
iii. E-mail Attachments Level

However, as time is a major constraint factor, the scope of this study will be limited to focusing on implementing the protection at network level. Although the focus is small, but it shall meet its objective as filtering at network level will reduce spam in the server, thus improving accuracy of spam blocking. It will also eliminate extra load in the server when number of e-mail to process is getting less.

For the mentioned test, a collection of spam will be fed into the mentioned antispam system, where these samples are actually collected from actual spam. However, the samples are limited to spam written in English characters only and the image spam will not the scanned for optical character recognition (OCR).

1.7 Hardware Specification for simulation

An existing computer with the following specification will be used to develop the mentioned antispam system:

Table 1.1: Specification of the Computer for the Simulation.

No.	Type	Specifications
5.	CPU	AMD Phenom (Quad Core)
6.	RAM	2 GB
7.	Hard Disk	500GB
8.	Operating System	OpenSolaris

A justification on choosing a unix machine will be in Chapter 3.

1.8 Summary

Problems as described in this chapter have shown that there is a need to improve the existing techniques and methods for a better result includes improving the utilization of system resources. As the problems are narrowing into catch rate and performance issues, therefore this chapter focuses on three objectives in resolving the issues. Firstly, study on other solution. Second objective is to form a pattern-based filtering technique based on combination of solutions. Third, to developed an antispam solution containing the said solution and perform a performance evaluation for a comparison with Antispam with Bayesian filter.

REFERENCES

Bekman, S. (2006). Anti-SPAM Techniques: Bayesian Content Filtering. Retrieved on Mayh 15th 2006, from http://stason.org/articles/technology/e-mail/junk-mail/bayesian_content_filtering.html

Foxman, E.R, and Schiano, W.T. (2000). Inspecting Spam Unsolicited Communications on the Internet. *Challenges of Information Technology Management in the 21st Century: 2000 Information Resources Management association International Conference*. May 21-24, 2000. Anchorage, Alaska. 552.

Karlberger, C., Bayler, G., Kruegel, C., and Kirda, E. (2007). Exploting Redundancy in Natural Language to Penetrate Bayesian Spam Filters. *FWF Austrian Science Fund*. 2007.

Kim, H.J, Kim, H.N, Jung, J.J, and Jo, G.S (2004). Spam Mail Filtering System Using Semantic Enrichment. WISE 2004, LNCS 3306. November 22-24, 2004. Brisbane, Australia. 619-628.

Liang, G., Li, T., Gong, X., Jiang, Y., Yang, J., and Ni, J. (2006). NASC: A Novel Approach for Spam Classification. *ICIC 2006, LNBI 4115*. August 16-19, 2006. Kunming, China. 672-681.

Mehta, B., Nangia, S., Gupta, M., Nejdl, W. (2008). Detecting Image Spam using Visual Features and Near Duplicate Detection. International World Wide Web Conference Committee (IW3C2). April 21-25 2008, Beijing China. 497-506.

McDonald, A. (2004). *SpamAssassin: A Practical Guide to Integration and Configuration*. Packt Publishing Ltd.

Nhung, N.P., and Phuong, T.M. (2007). An Efficient Method for Filtering Image-Based Spam E-mail. *Computer Analysis of Images and Patterns: 12th International Conference, CAIP 2007*. August 27-29, 2007. Vienna, Autria. 945-953.

Prakash, V.V., and O'Donnel, A. (2005). Fighting Spam with Reputation Systems. *Social Computing*. Vol. 3 (Issue No 9). ACM.

Ramachandran, A. and Feamster, N. (2006). Understanding the Network-Level Behaviour of Spammers. SIGCOMM'06. September 11-15 2006, Pisa Italy. 291-302.

Schwartz, A. (2004). *SpamAssassin*. California. O'Reily.

Schultz, B. (2004). TurnTide Stopping Spammers at Their Own Servers. *Network World*. 2004. Retrieved on April, 26, 2004. From http://www.networkworld.com/nw200/2004/0426watch9.html

Xiaochun Cheng, Xiaoqi Ma, Long Wang, and Shaochun Zhong (2005). A Mobile Agent Based Spam Filter System. *Computational Intelligence and Security: International Conference, CIS 2005*. December 15-19, 2005. Xi'an, China. 422-427.

Zdziarski, J.A. (2005). *Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification*. San Francisco. No Starch Press, Inc.

.