SPAM FILTERING USING ML ALGORITHMS

Md. Rafiqul Islam
School of Information Technology, Deakin University, Victoria, Australia.

Morshed U. Chowdhury
School of Information Technology, Deakin University, Victoria, Australia.

ABSTRACT

Spam is commonly defined as unsolicited email messages, and the goal of spam categorization is to distinguish between spam and legitimate email messages. Spam used to be considered a mere nuisance, but due to the abundant amounts of spam being sent today, it has progressed from being a nuisance to becoming a major problem. Spam filtering is able to control the problem in a variety of ways. Many researches in spam filtering has been centred on the more sophisticated classifier-related issues. Currently, machine learning for spam classification is an important research issue at present. Support Vector Machines (SVMs) are a new learning method and achieve substantial improvements over the currently preferred methods, and behave robustly whilst tackling a variety of different learning tasks. Due to its high dimensional input, fewer irrelevant features and high accuracy, the SVMs are more important to researchers for categorizing spam. This paper explores and identifies the use of different learning algorithms for classifying spam and legitimate messages from e-mail. A comparative analysis among the filtering techniques has also been presented in this paper.

KEY WORDS

Spam, SVM, Kernel functions, Machine Learning (ML), VC dimension.

1. INTRODUCTION

The Internet is gradually becoming an integral part of everyday life. Internet usage is expected to continue growing and e-mail has become a powerful tool intended for idea and information exchange, as well as for users' commercial and social lives. Along with the growth of the Internet and e-mail, there has been a dramatic growth in spam in recent years [16]. The majority of spam solutions deal with the flood of spam. However, it is amazing that despite the increasing development of anti-spam services and technologies, the number of spam messages continues to increase rapidly.

The increasing volume of spam has become a serious threat not only to the Internet, but also to society. For the business and educational environment, spam has become a security issue. Spam has gone from just being annoying to being expensive and risky. The enigma is that spam is difficult to define. What is spam to one person is not necessarily spam to another. Fortunately or unfortunately, spam is here to stay and destined to increase its impact around the world. It has become an issue that can no longer be ignored; an issue that needs to be addressed in a multi-layered approach: at the source, on the network, and with the end-user.

Consequently, spam filtering is able to control the problem in a variety of ways. Identification and spam removal from the e-mail delivery system allows end-users to regain a useful means of communication. Many researches on spam filtering have been centred on the more sophisticated classifier-related issues. Currently, machine learning for spam classification is an important research issue at present. The success of machine learning techniques in text categorization has led researchers to explore learning algorithms in spam filtering [1,12,18]. In particular, Bayesian techniques, support vector machines (SVM) effectively used for text categorization which influences researchers to classify the email is based on a special case of TC (text categorization), with the categories being spam and non-spam.

This paper explores and identifies the use of different anti spam techniques as well as statistical learning algorithms such as support vector machine (SVM) for classifying spam. A comparative analysis of various filtering techniques has been presented in this paper. A support vector machine is a new learning algorithm which has some attractive features, such as eliminating the need for feature selections which makes for easier spam classification. The organization of this paper is as follows: Section 2 describes the overview of spam, and problems associated with spam. In section 3, different techniques for spam filtering has been explored and in section 4 support vector machine has been described as spam filtering. Finally, a comparative analysis among the filtering techniques is presented in section 5 and the paper ends with conclusion and references.

2. SPAM PROBLEMS

Spam is difficult to define. In fact, there is no widely agreed or clear workable definition at present. We all recognize spam when we see it, but the truth is that what is spam to one person may not be spam to another. So, the notion of spam is subjective [19]. The most well-known definition seems to be 'unsolicited commercial email' (UCE) and 'Unsolicited Bulk Email' (UBE). The content of spam ranges enormously from advertisements for goods and services to pornographic material, financial advertisements, information on illegal copies of software, fraudulent advertisements and/or fraudulent attempts to solicit money.

More recently, spam has been spreading at an increasingly rapid rate, and while groups of spammers were relatively small in the past, the wide availability of 'spam kits' over the Internet has spread the practice from the United States to China, Russia and South America [13,15]. The scale of the problem is perhaps best highlighted when the growth of spam since 2001 is considered, as well as the percentage of spam, which was 7 per cent of all received e-mail [14]. By 2002 this had grown to 29 per cent, and by the end of 2003 the total stood at 54 per cent [16]. In March 2004 the percentage had increased to 63 per cent and this is set to continue rising. According to Message Labs, a US consultancy firm, spam now accounts for around 65% of all e-mail traffic. The following subsections outline a number of reasons which can explain why spam has become a serious problem.

2.1 Problems Related to Costs

Spam imposes costs on all Internet users. These costs have been increasing with the growth of the number of spam messages infiltrating the Internet daily. It is difficult to calculate the total costs of spam at the global level, though estimates suggest the costs are high. For example, a European Union (EU) study estimates that the worldwide cost of spam to Internet subscribers is in the vicinity of EUR 10 billion per year [14]. A June 2003 report [19] predicts that e-mail spam will cost companies USD 20.5 billion in 2003, and nearly ten times that amount, USD 198 billion, by 2007.

2.2 Problems Related to Privacy

The main privacy problem is that it causes significant unwanted intrusions. In addition, the collection of e-mail addresses is frequently made without the users' knowledge, much less with a specification of the purpose and consent. These problems are exacerbated when spam is sent indiscriminately.

2.3 Problems Related to Spam Content

The content of spam messages may create a problem due to fraud and deception. Fraudulent or deceptive spam can exist in a number of forms. Spammers disguise the origin of their messages because they know their messages are being blocked or filtered and they aim to entice individuals to open their messages A common trick that spammers use is to forge the headers of messages. Spammers often use the relay function in mail servers managed by others [15]. Some spam messages contain pornographic photographs and promote adult entertainment products and services.

2.4 Security Implications

Spam can temporarily or even permanently damage personal computers as well as clog computer networks. Large volumes of spam can interfere with critical computer infrastructures and endanger public safety. Spam may also be used maliciously as a Denial of Service (DoS) attack [14]. Some spam also contains destructive viruses and worms. According to estimation [16], 90% of viruses are passed through e-mail and 51% of corporations have had a virus disaster and /or computer worms. The experience of spam linked with viruses has led to a greater mistrust of e-mail as a secure communication mechanism.

2.5 Identity Theft

Identity theft is on the rise, threatening e-commerce by eroding consumer trust. Every e-mail contains information regarding its origin, but current technology does not guarantee that the information on the header is correct. If spammers discover that all e-mail from a particular company is allowed through spam filters because the company is on a white-list, spammers can make their e-mails look like they originate from that source. Spammers usually use some other business' IP address or conceal their own identity by using stolen or falsely labelled company identities [16]. Others alter the header to falsify the sender or create an open relay through unsecured servers.

3. SPAM FILTERING

3.1 Rule Based Spam Filtering

The first automated filtering techniques to be discussed are those which use a set of rules to classify e-mail as spam or legitimate e-mail. The rule based filtering techniques can be applied at either the MUA (Mail User Agent) level or the MTA (Mail Transfer Agent) level.

3.1.1 MUA Rule Based Filters

E-mail clients contain an element at the MUA level for categorising e-mail based on a set of rules determined by the user. These rules can be constructed to examine an e-mail message's body, for keywords or phrases given by the end-user. A common use of such rules is to categorise newly arrived e-mail into a specific folder. This MUA based model is illustrated in Figure 1.

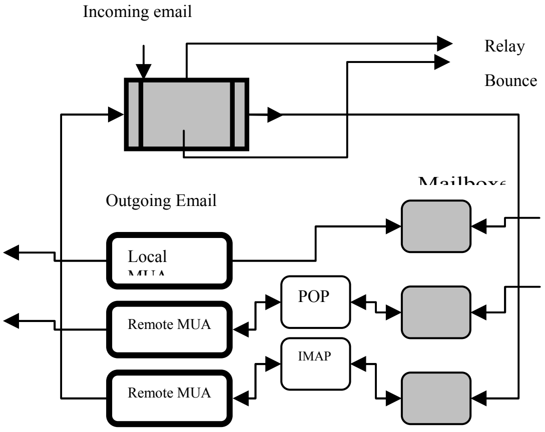

Figure 1. The MUA based model

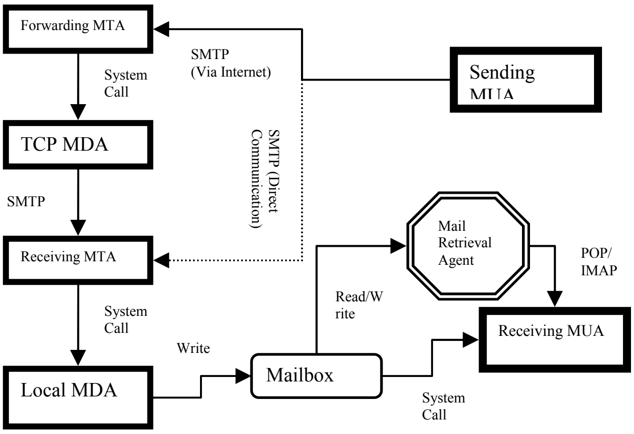

Figure 2. MTA based Mail Transferring Mechanism

The user could create a folder called spam and define a number rules that would transfer a newly arrived e-mail to the spam folder if it were triggered. Such rules could look for specific words in the content of the e-mail, look for punctuation being used in the subject of the e-mail, or note the content type of the e-mail.

While this technique does work well, it does have a serious problem. The rule set needs constant updating and refinement because most spammers use obfuscation techniques. Some common obfuscation used is misspelling words.

3.1.2 White-list/Blacklist

Whitelist is an MUA level rule-based filtering technique, where a whitelist is a register containing a collection of contacts from which e-mail messages can be accepted. If an e-mail arrives but does not come from one of the contacts in the whitelist, then it is treated as spam and placed in the spam folder. While this technique is effective for some users, it has also drawbacks. Any email sent by a stranger will simply be incorrectly classified as false positive (FP). However there is a scheme that incorporates a challenge response mechanism to allow users to be added to a user's whitelist.

A blacklist contains lists of known spammers. Essentially when a user gets spam, the user adds the sender of the spam to the blacklist. The entire domain of the sender of the spam can be added to the blacklist. Newly arrived e-mails are checked, and if the sender is on the blacklist, the e-mail is automatically classified as spam. As with the whitelist, there are flaws with blacklists too. The major problem stems from the fact that spammers tend to forge header information in their spam. The sender information is generally forged, meaning that perhaps innocent people are added to a blacklist but more importantly the effect which the blacklist will have, is diminished dramatically.

3.1.3 MTA Rule Based Filtering

Filtering at the MTA level can achieve some economies of scale but it also triggers some problems. Since by nature, spam is sent in bulk, blocking the sender can dramatically reduce the number of spam needed to be stored and delivered. Some of the techniques described for MUA rule based filtering can be applied at the MTA level [15]. Figure 2 illustrates the general email transferring mechanism using MTA.

3.1 4 Distributed Blacklist

A distributed blacklist is a network tool for anti-spam engines. Distributed blacklists maintain a collection of common spam messages on a central server. The filter is shared amongst the subscribers, so if one person identifies a message as spam then all others benefit. When a message arrives, it is compared to the digest of known spam and deleted if a match is found. This method is low in false-positives, but false-negatives tend to be high so often another filtering technique is required to work in conjunction. The central repository must be maintained by an unbiased organisation [15,13].

3.2 Content Based Spam Filtering

Spam will typically have a distinctive content, which should be easy to distinguish from legitimate e-mail. Categorising e-mail based on its content seems like a logical progression from simplistic rule based approaches. This would help reduce error rates as legitimate e-mail would not be blocked even if the ISP from which it originated, is on a real-time block list. In addition, the presence of a single token should not cause the e-mail to be classified as spam.

In content hash based spam filtering, when a new e-mail arrives its hash/messages-digest can be compared to a list of know spam hashes. If there is a match then the e-mail can be deleted without fear. The true strength of this technique only becomes apparent when we introduce a distribution mechanism for known spam hashes [19]. In order to circumvent the hash based filters, spammers started to introduce some random variables into spam. The result was that each and every spam now had a unique hash value. When the hash of a spam was computed, its hash would not match that of any known spam since it contains some random variables. Due to the strict avalanche affect, the output was completely different for each spam which has some random content [9].

3.3 Personalised, Collaborative Spam Filtering

Research has been conducted into the CASSANDRA architecture for building personalised, collaborative spam filters [15]. In this system, users act as peers in an adaptive P2P network. Each time a spam is identified as such by a peer, a spam notice is generated and sent to the peers most likely to receive a similar spam. In order to determine which peers are "most like" [15] a given peer, each peer maintains a history of previous interactions with others. If two peers have shown themselves to receive similar spam and have not generated any information that causes false positives, then they will cluster towards each other in the network and maintain connections. When a peer receives a new, previously unidentified spam, it notifies the peers to whom it has connections.

The key advantage of this system is its resilience and adaptability. Spam has been shown to exhibit "concept drift" [9], which is the change in the characteristic content of spam over time. A disadvantage of this approach is that it uses SMTP to communicate between peers. SMTP is unauthenticated, and so malicious attackers could attempt to undermine the filter by spoofing reports from other peers. This can be addressed by using the CTK [19].

4. MACHINE LEARNING FOR SPAM FILTERING

Spam filtering based on the textual content of email messages can be seen as a special case of text categorization, with the categories being spam and non-spam. Although the task of text categorization has been researched extensively, its particular application to email data and detection of spam specifically is relatively recent. Some initial research studies [13,17] primarily focused on the problem of filtering spam whereby Naïve Bayes (NB) was applied to address the problem of building a personal spam filter. NB was advocated due to its previously demonstrated robustness in the text-classification domain and due to its ability to be easily implemented in a cost-sensitive decision framework. Although high performance levels were achieved using word features only, it was observed that by additionally incorporating non-textual features and some domain knowledge, the filtering performance could be improved significantly. The application of SVMs to the spam-filtering task was also suggested.

4.1 Support Vector Machines (SVMs) for Spam Filtering

The SVM is a classification and regression algorithm which was developed by Vapnik [19] and it is gaining popularity due to many attractive features, and its promising empirical performance. SVMs contain a range of classification and regression algorithms that have been based on the Structural Risk Minimization (SRM) principle from statistical learning theory formulated by Vapnik [7, 9,10,19]. The role of the SRM is to find an optimal hyperplane for which the lowest true error can be guaranteed. This framework has developed into a learning algorithm when trained from a finite data set, and formed the 'true' performance when used in practice. The key concepts of SVMs can be categorised into two classes, $y_i \in \{-1,1\}$, and there are n labelled training examples $\{x_1, y_1),\ldots,(x_n, y_n), x \in R^d$ where d is the dimensionality of the vector.

SVMs are based on the idea that every solvable classification problem can be transformed into a linearly separable one by mapping the original vector space into a new one, using non-linear mapping functions. More formally, SVMs learn generalized linear discriminant functions of the following form:

$$f(\vec{x}) = \sum_{i=1}^{m'} w_i . h_i(\vec{x}) + w_0$$

where m′ is the dimensionality of the new vector space, and $h_i(\vec{x})$ are the non-linear functions that map the original attributes to the new ones. The higher the order of the $h_i(\vec{x})$ functions, the less linear the resulting discriminant. The type of $h_i(\vec{x})$functions that can be used is limited indirectly by the algorithm's search method, but the exact choice is made by the person who configures the learner for a particular application. The function $f(\vec{x})$ is not linear in the original vector space, but it is linear in the transformed one.

The search method of an SVM aims to select the hyperplane that separates the training instances (messages) of the two categories with maximum distance (Figure 3 & 4). This target hyperplane is found by selecting two parallel hyperplanes that are each tangential to a different category - i.e., they include at least one training instance of a different category, whilst providing perfect separation between all the training instances of the two categories. The training instances that lie on, and thus define the two tangential hyperplanes are the support vectors. The distance between the two tangential hyperplanes is the margin. Once the margin has been maximized, the target hyperplane is in the middle of the margin.

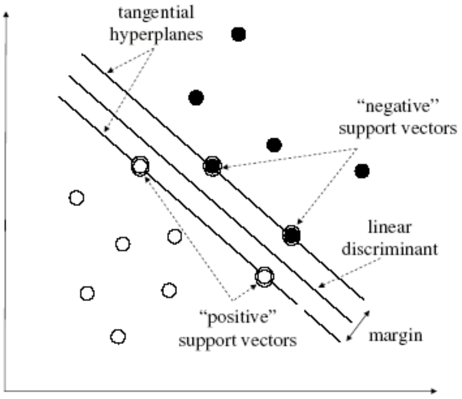

Figure 3. Linear discrimination with SVMs in a linearly separable case

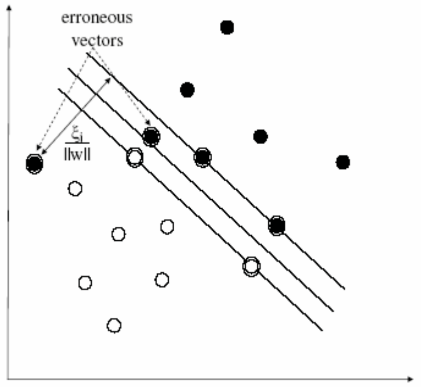

Figure 4. Linear discrimination with SVMs in a case that is not linearly separable

4.2 Support Vector Kernel

The basic idea of a kernel is that it gives the equivalent of mapping a nonlinear separable input space, to a higher dimensional feature space that is linearly separable. An important concept behind kernels is how to simplify the classification task of nonlinearly separating data. One way is to find a function φ (Figure 5) that maps the input space X to some feature space F where the problem is linearly separable, thus classifying the data in the new feature space. The problem here is the new feature space can have a very high number of dimensions which can make the computation of the classification task unfeasible [20].

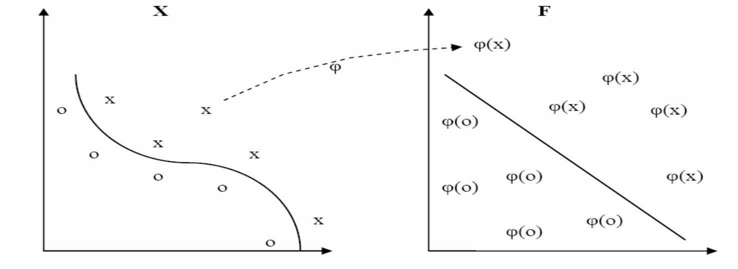

Figure 5. Simplify the Classification task.

The choice of kernel affects the model bias of the algorithm. The training vectors $\vec{x}$ are mapped onto a higher (maybe infinite) dimensional space by the function Φ . Then an SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. C > 0 is the penalty parameter of the error term. Furthermore, $K(\vec{x}_j, \vec{x}_k) = \Phi(\vec{x}_j)^T \Phi(\vec{x}_k)$ is called the kernel function. Though new kernels are being proposed by researchers, the four types of kernel functions frequently used with SVM:

Linear: $k(x_i, x_j) = x_i^T x_j$;

Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$;

RBF: $K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0$; and

Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$, here, γ ,*r*, and *d* are kernel parameters.

The obvious dilemma that arises is that, it is difficult to decide on which kernel is best to select for a particular problem. This is a critical situation however it is easier to make a comparison with the inclusion of many mappings within one framework. So far, kernel are used for getting high dimensional feature space are trail and error basis. There is no specific technique to detect which kernel is best for a particular problem.

5. A COMPARATIVE ANALYSIS AMONG FILTERING TECHNIQUES

The following Table 1 summarizes the compare and contrast of different spam filtering techniques.

Table 1. Benefits and limitations of spam fingering methods

Techniques	Benefits		Limitations
Rule Based Spam Filtering	MTA/ MUA	It is easy to install and effective in blocking a large percentage of spam	• The rule set needs constant updating and refinement because most spammers use obfuscation techniques. • It should be combined with other methods to filter out a larger volume of spam
	Blacklist/ Distributed Black list	• Blocks mail from known spam sources • Readily available pools of lists • It is effective and easy to implement	• May block harmless messages • Needs constant updating and maintenance • Exact rules are difficult to formulate and maintain-Spam is always changing spammer e-mail addresses • Any email sent by a stranger will simply be incorrectly classified as FP/FN
	Whitelist	Guarantees delivery of known good addresses	List maintenance can be cumbersome
Content Based Filtering	• It reduces error rates as legitimate e-mail would not be blocked even if the ISP from which it originated, is on a real-time block list. • The presence of a single token should not cause the e-mail to be classified as spam.		Need occasional refinement and in most cases the refinement is automated, meaning less hassle for end-users.
Content Hash Based Filtering	• The true strength of this technique only becomes apparent when we introduce a distribution mechanism for known spam hashes. • Blocks known spam • Low rate of false positives		• Spammers introduce some random variables/characters either in the body or on the subject line. So, this would create completely different outputs for each spam which has some random content. • Time-sensitive • Can be circumvented by randomization
Personalised, Collaborative Filtering	• The key advantage of this system is its resilience and adaptability, because personalised collaborative spam filters continually refine their contacts with whom they are connected. • In addition, possible eradication of large volumes of spam through collaborative reporting of spam		• A disadvantage of this approach is that it uses SMTP to communicate between peers. SMTP is unauthenticated, and so malicious attackers could attempt to undermine the filter by spoofing reports from other peers. • Still vulnerable to random changes in spam e-mail, and there are problems with scalability of this method
Machine learning techniques	• It is very effective and is also adaptive, so hard to fool • Based on text classification methods: TF-IDF, Naïve Bayes, N-gram, SVM, Boosting etc. • Phenomenally accurate • Learns new spammer tactics automatically • Adapt to changing spam		• Need lots of training data • Spammers are learning too-Images, synonyms, misspellings, … • Hard to get good email corpora • Need huge attributes. • Functions best with individual user settings • Accuracy dramatically decreases when deployed as a generic gateway solution • Requires more processing power
Support Vector	• High dimensional feature Space • Can handle more than 30,000 attributes		• Difficult for non-linear separable case • Training time is high compare to NB

Machines	• Kernel based learning algorithm • Accuracy is high at classification time	

6. CONCLUSION

In this paper, a thorough investigation of spam and spam filtering using different techniques has been presented as well as detailing spam problems. Emphasis was based on different aspects of anti-spam filter, especially the learning-based anti-spam filter. A comparative study among the different anti-spam techniques has also been presented. In addition, support vector machine based spam filtering techniques has discussed because SVM has some attractive features, such as eliminating the need for feature selections which makes for easier text categorization, which are more important for spam filtering.

REFERENCES

Androutsopoulos, I. et. al,. 2004, Learning to Filter Unsolicited Commercial Email. *NCRS, T. Report*.

Burges, C. 1998. A tutorial on Support Vector machines for pattern recognition. *Journal of data Mining and Knowledge Discovery* 2, 2, 121–167.

Brightmail 2004, The Brightmail website. Online, available: www.brightmail.com/spamstats.html.

Cortes C., and Vapnik V, 1995, "Support vector networks," *Mach. Learn.* 20 (1995) 273–297.

Cohen, W. 1996. Learning rules that classify e-mail. *In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access.* Palo Alto, California, 18–25.

Cranor, L. and LaMacchia, B. 1998. Spam! Communications of the *ACM* 41, 8, 74–83.

Cristianini, N. and Shawe-Taylor, J. 2000. An introduction to Support Vector Machines and other kernel-based learning methods. *Cambridge University Press*.

CyberAtlas staff "Population explosion", March 14, 2003, Online, available: http://cyberatlas.internet.com /big_picture/geographics/article/0,,5911_151151,00.html

Cunningham, P., N. Nowlan, S. J. Delany, and M. Haahr, 2003 "A Case-Based Approach to Spam Filtering that Can Track Concept Drift", in *Proceedings of the ICCBR'03 Workshop on Long-Lived CBR Systems.*

Drucker, H. D., Wu, D., and V., V. 1999. Support Vector Machines for spam categorization. *IEEE Transactions On Neural Networks* 10, 5, 1048–1054.

Gray .A and M. Haar, 2004 "Personalised, Collaborative Spam Filtering", in *Proceedings of the FirstConference on Email and Anti-Spam (CEAS)*, 2004.

Joachims, T. 1998, Text categorization with Support Vector Machines: Learning with many relevant features. *In Proceedings of the 10th European Conference on Machine Learning, Lecture Notes in Computer Science*, n. 1398. Springer Verlag, Heidelberg, Germany, 137–142.

Linford, S. 2004, Interview with Steve Linford (Spamhaus) conducted on 9th January, 2004.

Self, K.M., 2004 "Challenge-Response Anti-Spam Systems Considered Harmful", Website, 2004, http://kmself.home.netcom.com/Rants/challenge-response.html.

Thomson, I., 2003. Mafia muscles in on spam and viruses. Vnunet.com. Online, Available: http://www.vnunet.com/News/1151421.

Salem,E. 2004, Interview with Enrique Salem (CEO of Brightmail) conducted on 23rd Feb, 2004.

Sahami M, et. al., 1998, A bayesian approach to filtering junk e-mail. In Learning for Text Categorization: *Papers from the 1998 Workshop*, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05.

Sebastiani, F. 2002 Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1, pp.1–47.

Seigneur J-M, and Jensen C D, 2004, "The Claim Tool Kit for Ad-hoc Recognition of Peer Entities", in *Journal of Science of Computer Programming*, Elsevier.

Zhang, J., et. al,. A. 2003, Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization. *In Proceedings of the 20th International Conference on Machine Learning. AAAI Press*, pp.888–895.