International Journal of the Physical Sciences Vol. 5(12), pp. 1869-1882, 4 October, 2010
Available online at http://www.academicjournals.org/IJPS
ISSN 1992 - 1950 ©2010 Academic Journals

Full Length Research Paper

Overview of textual anti-spam filtering techniques

Thamarai Subramaniam, Hamid A. Jalab and Alaa Y. Taqa*

Computer System and Technology, Faulty of Computer Science and Information Technology, University Malaya, Malaysia.

Accepted 31 August, 2010

Elecronic mail (E-mail) is an essential communication tool that has been greatly abused by spammers to disseminate unwanted information (messages) and spread malicious contents to Internet users. Current Internet technologies further accelerated the distribution of spam. Effective controls need to be deployed to countermeasure the ever growing spam problem. Machine learning provides better protective mechanisms that are able to control spam. This paper summarizes most common techniques used for anti-spam filtering by analyzing the e-mail content and also looks into machine learning algorithms such as Naïve Bayesian, support vector machine and neural network that have been adopted to detect and control spam. Each machine learning has its own strengths and limitations as such appropriate preprocessing need to be carefully considered to increase the effectiveness of any given machine learning.

Key words: Anti-spam filters, text categorization, electronic mail (E-mail), machine learning.

INTRODUCTION

E-mail or electronic mail is an electronic messaging system that transmits messages across computer networks. Users simply type in the message, add the recipient's e-mail address (es) and click the send button. Users can access any free e-mail service such as Yahoo mail, Gmail, Hotmail, or register with ISPs (Internet Service Providers) in order to obtain an e-mail account at no cost except for the Internet connection charges. Besides that, e-mail can be also received almost immediately by the recipient once it is sent out.

E-mail allows users to communicate with each other at a low cost as well as provides an efficient mail delivery system. The reliability, user-friendliness and availability of a wide range of free e-mail services make it most popular and a preferred communication tool. As such, businesses and individual users alike rely heavily on this communication tool to share information and knowledge. Businesses can drastically cut down on communication cost since e-mail is extremely fast and inexpensive; furthermore it is a very powerful marketing tool. Businesses can capitalize from this technology since it is a very popular advertising tool. However, the simplicity of sending e-mail and the almost non-existent cost poses another problem: Spam. Spam refers to bulk unsolicited commercial e-mail sent indiscriminately to users. Table 1 enumerates some of them.

Based on the Ferris Research (2009), spam can be categorized into the following:

1. Health; such as fake pharmaceuticals;
2. Promotional products; such as fake fashion items (for example, watches);
3. Adult content; such as pornography and prostitution;
4, Financial and refinancing; such as stock kiting, tax solutions, loan packages;
5. Phishing and other fraud; such as "Nigerian 419" and "Spanish Prisoner";
6. Malware and viruses; Trojan horses attempting to infect your PC with malware;
7. Education; such as online diploma;
8. Marketing; such as direct marketing material, sexual enhancement products;
9. Political; US president votes.

E-MAIL STRUCTURE

E-mail messages are divided into 2 parts: Header

*Corresponding author. E-mail: alaa_taqa@um.edu.my.

Table 1. Different spam definitions.

Author(s)/(year)	Definition
Vapnik et al. (1999)	An e-mail message that is unwanted: Basically it is the electronic version of junk mail that is delivered by the postal service.
Oda and White (2003)	The electronic equivalent of junk e-mail which typically covers a range of unsolicited and undesired advertisements and bulk e-mail messages.
Lazzari et al. (2005)	Electronic messages posted blindly to thousands of recipients, and represent one of the most serious and urgent information overload problems.
Zhao and Zhang (2005)	Spam or junk mail, is an unauthorized intrusion into a virtual space - the E-mail box.
Youn and McLeod (2007)	Spam as bulk e-mail - e-mail that was not asked for which is send to multiple recipients.
Wu and Deng (2008)	Spam e-mails, also known as 'junk e-mails', are unsolicited ones sent in bulk (unsolicited bulk E-mail) with hidden or forged identity of the sender, address, and header information.
Amayri and Bouguil (2009)	Spam e-mails can be recognized either by content or delivery manner and indicated that spam e-mails were recognized according to the volume of dissemination and permissible delivery.
Spamhaus (2010)	An electronic message is "spam" if (A) the recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients; AND (B) the recipient has not verifiably granted deliberate, explicit, and still-revocable permission for it to be sent

information and message body. Header information or the header field consists of information about the message's transportation which generally shows the following information;

1. From: displays sender's detail such as e-mail address;
2. To: displays receiver's detail such as e-mail address;
3. Date: displays the date the e-mail was send to the recipient;
4. Received: intermediary server's information and the date the e-mail message is processed;
5. Reply to: reply address;
6. Subject: the subject of message specified by the sender;
6. Message Id: unique id of the message and others

The message body contains the message of the e-mail. E-mail messages are presented in plain text or HTML. An e-mail may also have attachments such as graphics, video or other format type and to facilitate these attachments MIME (multipurpose internet mail extension) is used.

SPAMMER TRICKS

In order to send spam, spammers first obtain e-mail addresses by harvesting addresses through the Internet using specialized software. This software systematically gathers e-mail addresses from discussion groups or websites (Schaub, 2002), other than that spammer also able to purchase or rent collections of e-mail addresses from other spammers or services providers. Table 2 indicates the many tricks used by spammers to avoid detection by spam filters.

SPAM's IMPACTS

The MessageLabs Intelligence report for 2009 highlights spam levels reaching 87.7%, with compromised computers issuing 83.4% of the 107 billion spam messages distributed globally per day on average (MessageLabs Intelligence Annual Security Report, 2009). Spam reaching users' inbox have been gradually increasing since 2004 as shown in Figure 1 (data is

Table 2. Tricks used by spammers to send spam.

Tricks	Descriptions
Zombies or Botnets	Compromised PCs on the Internet that sent vast amount of spam, viruses, and malware.
Bayesian sneaking and poisoning	Writing spam message so it does not contain any words that are normally used in spam messages, or "poison" the Bayesian filter's database.
IP address	Borrowing or using an IP address that has a good or neutral reputation.
Offshore ISPs	Usage of offshore ISPs that lack in security measures
Open proxies / open-relay servers	Compromised servers to re-direct spam to unsuspecting users.
Third-party mailback software	Use improperly-secured mailback applications on innocent websites
Falsified header information	Add bogus header information to the spam message
Obfuscation	Obscuring the words in spam messages by splitting words or messages using nonsense HTML tags or other 'creative' symbols
Vertical slicing	Writing the spam messages vertically
HTML manipulation	Manipulation of HTML format to avoid detection
HTML encoding	Usage of encoding scheme such as Base64 to turn a binary attachment into plain text characters
JavaScript messages	Placing entire contents of the spam message inside a JavaScript snippet that is activated when the message is opened
ASCII art	Usage of letter glyphs of standard letters to write spam messages
Image based	Using image to send textual information
URL address or redirect URL	Only add URL address to bypass detection / use expendable "portals" to point to their actual websites
Encrypted messages	Encrypting message where it only decrypted once it reaches the mail box

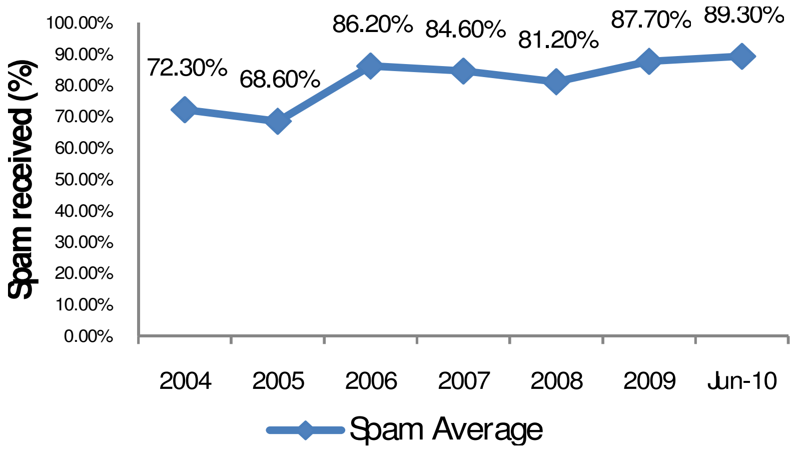

Figure 1. Spam average for 2004 to 2010.

compiled from MessageLabs Intelligence reports for 2005, 2006, 2007, 2008, 2009, 2010). The decline in the year 2005 is contributed due to awareness campaign launched on 2004 aimed to pressure internet service

Table 3. Anti-spam legislation environment (Moustakas et al., 2010).

Country	Legislation – Anti-spam statutes
Australia	Spam Act of 2003 Telecommunications Act of 1997 Australia Parts IVA, V, and VC of the Trade Practices Act of 1974
Canada	Personal Information Protection and Electronic Documents Act (PIPEDA) Competition Act. Charter of Rights Freedoms The Criminal Code and the Competition Act Canadian Code of Practice for Consumer Protection in E-Commerce
EU	Privacy and Electronic Communication Regulations 2003 (UK) Data Protection Act of 1998 (UK) Electronic Commerce Regulations of 2002 (all adapted from EC Directives, e.g. Directive on Privacy and Electronic Communications 2002/58/EC)
Japan	Law on regulation of Transmission of Specified Electronic Mail July, 2002 Specific Commercial Transactions Law, 2002
USA	CAN-SPAM ACT of 2003 Law enforced by Federal Trade Commission Section 5 of the Federal Trade Commission Act

providers and other internet bodies to take a responsible role in helping to clamp down on e-mail attacks (MessageLabs Intelligence Annual Security Report, 2005).

A 2009 study by Ferris Research estimated an increase in spam cost to a total of $130 billion worldwide. This represents a 30% increase from 2007 (Ferris Research, 2009). The study indicated that the main cost occurs due to:

1. Productivity loss from inspecting and deleting spam that gets missed by spam control products (false negatives),
2. Productivity loss from searching for legitimate e-mail deleted in error by spam control products (false positives)
3. Operations and helpdesk running costs (Ferris Research, 2009).

The impacts of spam are becoming far more serious than mere annoyances. Spam floods up users' inboxes thus making users spend unproductive hours in deleting these unwanted e-mails, causing displacement of critical or legitimate e-mails. Besides that, spam also causes the loss of internet performance and bandwidth due to increased payload on the network (Ferris Research, 2010) and it clogs up e-mail servers to the point where it sometimes crashes.

Spam increases the spread of malware and viruses which poses bigger threats to network security and personal privacy (Lai et al., 2009). Based on a MessageLabs research report, spam containing viruses for 2009 was 1 in 286.4 e-mails and more than 73.1 million malware infected e-mails, containing over 2,500 different malware strains, were blocked (Wood et al., 2010).

Spammers also deploy spam to gain personal information about the user for fraudulent proposes. Phishing activity related to identify theft and other internet related frauds (e.g. Nigeria 419) are becoming one of the major concerns for the Internet community. MessageLabs researchers indicated that the proportion of phishing attacks in e-mail traffic was 1 in 325.2 (0.31%) e-mails and estimated 161 billion e-mail phishing attacks were in circulation in 2009. The growing threats of spam definitely require drastic control measures.

EXISTING SOLUTIONS FOR SPAM

Traditionally there are many approaches available to control spam such as using sender domain check, content check, open relay prohibition and checking the IP address or domain names (Hideo, 2009). However, spammers easily overcome these simple measures with more sophisticated variants of spam to evade detection. The measures engaged to control spam are discussed below.

Legislation approaches

Guzella (2009) cited that economical impacts of spam have led some countries to adopt legislation. Many countries (Table 3) have enacted different laws and legislations to protect businesses and individuals alike against spam. Denmark enacted the Danish Marketing Practices Act, Data Protection Act and Danish Act on Internet domains (Frost and Udsen, 2006) that prohibit spammers from harvesting and sending spam e-mails.

In USA, CAN-SPAM Act for 2003 was enacted in December 2003. CAN-SPAM Act is an abbreviation for controlling the assault of non-solicited pornography and marketing. It places restrictions and regulations to control spammers activities. For example, it prohibits spammers from harvesting e-mail addresses and creating Botnets. Failure to comply with CAN-Spam Act can result in a monetary penalty of $16,000 per incident.

However the CAN-Spam Act does allow spammers to send unsolicited e-mail. McAfee Research reported on 2009 despite the six-year-old CAN-SPAM Act, spammers routinely abuse the law and continue to deliver spam (Wosotowsky and Winkler, 2009).

Black-list and white-list

Besides legislation, technological spam detection approaches have also been employed over the years. Earliest techniques used to block spam were whitelist and blacklist. This content-based technique recognizes words or patterns of a message which are defined either legitimate mail or spam.

Legitimate mails are listed in a whitelist and spam is listed in a blacklist. The e-mail message is then analyzed against the lists and legitimate e-mails are allowed while spam mails are blocked. Unfortunately, since the context of the e-mail is not taken into consideration, some legitimate e-mail may be blocked or blacklisted (Dalkilic et al., 2009; Heron, 2009).

Messages from previously known source of spam are blocked using Real-time IP blacklist. Real-time IP blacklist typically checks the source of the spam. The header information from the messages which contain IP or domain sources is compared against real-time blacklist and matched IP addresses are blocked.

On the other hand spammers are using large Botnets to sent spam thus creating extremely a huge number of IP addresses to be blacklisted. Real-time IP blacklist typically blocks only 80 - 90% of spam (Green, 2005). Sometime the filter application blocks legitimate users (false positive) who have unknowingly been used to generate spam or have been erroneously reported (Heron, 2009). The time and effort it takes to remove these false positive can be overwhelming.

Heuristic approaches

Another approach used to control spam is heuristics. The heuristic approach examines the e-mail's content and compares it against thousands of pre-defined rules. These rules are assigned a numerical score that weight the probability of the message being spam. Each received message is verified against the heuristic filtering rules.

Compared with a pre-defined threshold, the verification result decides whether the message is spam or not (Xie et al., 2006). The score of the weight is then shared among users to filter the e-mails. Conversely, spammers use obfuscation to fool the rules to avoid detection and modifying heuristic tests to cope with new attack vectors devised by spammers which can be complicated, leaving a period of time when there is no protection (Heron, 2009).

Machine learning approaches

Machine learning (ML) is a scientific discipline that is concerned with the design and development of algorithms that allow computers to adapt their behavior based on data. ML automatically learns to recognize complex patterns and makes intelligent decisions based on data.

ML is capable of automatically building a classifier for a category by observing the characteristics of a set of documents or corpus manually classified under C_i by a domain expert. From these characteristics, the inductive process gleans the characteristics that a new, unseen document should have in order to be classified under C_i (Sebastiani, 2002).

ML's automatic builder of classifiers (learner) deems to be the main advantage when it comes to spam classification. It is more convenient and easier to automatically classify a set of documents than to build and tune a set of rules.

NAÏVE BAYESIAN CLASSIFICATION

Naïve Bayesian is a fundamental statistical approach based on probability initially proposed by Sahami et al. (1998). The Bayesian algorithm predicts the classification of new e-mail by identifying an e-mail as spam or legitimate. This is achieved by looking at the features using a 'training set' which has already been pre-classified correctly and then checking whether a particular word appears in the e-mail. High probability indicates the new e-mail as spam e-mail.

Lai (2007) describes naïve Bayesian algorithm as follows: Given a feature vector $x = \{x_1, x_2, x_3, \cdots, x_n\}$ of an e-mail, where values of attributes $x_1, x_2, x_3, \cdots, x_n$ and n is the number of attributes in the corpus. Each attribute is a particular word occurring or not in an e-mail. Let c denote the category to be predicted, that is,

$c \in \{spam, legitimate\}$, by Bayes law the probability that x belongs to c is as given in

$$P(c|x) = \frac{P(c)\cdot P(x|c)}{P(x)} \qquad (1)$$

where $P(x)$ denotes the a-priori probability of a randomly picked e-mail has vector x as its representation, P(c) is also the a-prior probability of class c (that is, the probability that a randomly picked e-mail is from that class), and $P(x|c)$ denotes the probability of a randomly picked e-mail with class c has x as its representation.

Androutsopoulos et al. noted that the probability $P(x|c)$ is almost impossible to calculate because the fact that the number of possible vectors x is too high. In order to alleviate this problem, it is common to make the assumption that the components of the vector x are independent in the class. Thus, $P(x|c)$ can be decomposed to

$$P(x|c) = \prod_{i=1}^{n} P(x_i|c) \qquad (2)$$

So, using the NB classifier for spam filtering it can be computed as

$$C_{NB} = \arg max_{c\in\{spam,legitimate\}} P(c) \prod_i P(x_i|c) \qquad (3)$$

Naïve Bayesian approach is very stable, better and has faster performance thus making it very popular (Dong, 2004) algorithm to employ in various classification fields. NB performs reasonably consistently and is good in different experimental settings (Lai, 2007). It is simple to implement and independence allows parameters to be estimated on different data sets. Besides that, NB also has a very short learning curve (Ko et al., 2009). The main shortcoming of the NB classifier is it can only learn linear discriminant functions and thus it is always suboptimal for non-linearly separable concepts (Rish, 2001). The Naïve Bayesian approach has been successfully incorporated into other machine learning approaches to increase the effectiveness of the text classifications.

SUPPORT VECTOR MACHINE CLASSIFICATION

Support vector machine (SVM) is a framework of structural risk minimization and statistical learning theory developed by Vapnik and his coworkers. SVM is based on the optimal classification hyperplane of linear classification situation. SVM finds a maximum margin separating hyperplane between two classes of data (Figure 2). It is a non-linear function and density estimation based algorithm.

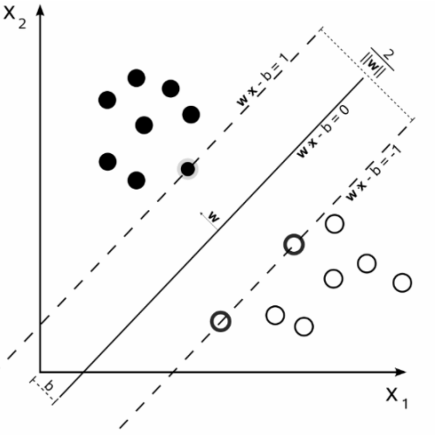

Figure 2. Support vector machine.

Sun et al. (2002) indicated that classifier built on SVM has shown promising results with its efficiency and effectiveness. SVM can be achieved by non-linear mapping, polynomial functions and sigmoidal functions. SVM bypasses the curse of dimensionality by employing kernel functions and enables the straight forward analysis of high-dimensional data, the ability to determine the margin completely as well as the capability of handling high dimensionality and small sample problems (Yu et al., 2008). SVM has a great generalization capability too (Sebastiani, 2002).

In SVMs hyperplane that separate the training, data (spam or legitimate e-mail) are measured by the maximum margin, therefore all vectors that lie on one side of the hyperplane are labeled as -1 (w x – b = -1) and the other side as +1 (w x – b = 1). Thus when new data is introduced, it maps to the closest support vector based on the maximum margin. To find the maximum margin, the following algorithms are used, given linear separable vectors $x_i\ (\mathrm{i}= 1,\cdots,l), x_i \in R^n$ with labels $y_i = \pm 1$, and for linearly separable space, the decision surface is a hyperplane which can be written as:

$$w * x + b = 0 \qquad (4)$$

And the equation is

$$\text{MIN } g(w) = \frac{1}{2}||w||_2^2 \qquad (5)$$

With a constraint

$$y_i(w * x_i + b) \geq 1, i = 1,\cdots,l \qquad (6)$$

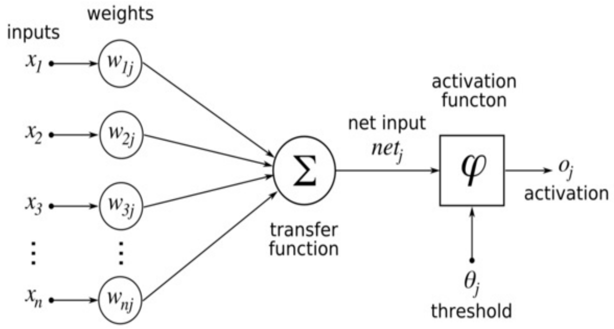

Figure 3. Neural network: Indicates the input, hidden and output layer that make up the neural network.

The optimal hyperplane calculation as follows

$$\sum_{i=1}^{l} y_i \alpha_i (x_i * x) + b_0 = 0 \tag{7}$$

NEURAL NETWORK CLASSIFICATION

Neural network (NN) was first introduced by McCulloch and Pitts in 1943, since the introduction it has been increasingly used in text classification. Neural network emulates the functionality of human brains in which neurons (nerves cell) communicate with each other by sending messages between them. Artificial neural network (ANN) represents the mathematical model of these biological neurons.

It is a parallel distributed information processing structure consisting of a number of nonlinear processing units (neurons) (Ko et al., 2009) which can be trained to recognize features and to identify incomplete features/ data. Neural network has great mapping capabilities or pattern association thus exhibiting generalization, robustness, high fault tolerance, and high speed parallel information processing.

NN's self-learning capability by examples allows researchers to train NN with features from e-mail messages to acquire the knowledge for classifying e-mail into spam or legitimate mail. Neural network architecture generally can be categorized into single layer feedforward network, multilayer feedforward network and recurrent network. However over the years many other types have emerge such as perceptron, back-propagation network, self-organizing map, adaptive resonance theory and radial basis function.

Figure 3 indicates the input, hidden and output layer that make up the neural network.

The network functions are as follows (Goyal, 2007): Each node i in the input layer receives a signal x_i as the network's input, multiplied by a weight value between the input layer and the hidden layer. Each node j in the hidden layer receives the signal $In(j)$ according to the following equation.

$$In(j) = \theta_j + \sum_{i=1}^{n} x_i w_{ij} \tag{8}$$

This is then passed through to the bipolar sigmoid activation function

$$f(x) = \frac{2}{(1+\exp(-x))} - 1 \tag{9}$$

The output of the activation function $f(In(j))$is then broadcast to all of the nodes on the output layer

$$y_k = \theta_k + \sum_{j=1}^{m} w_{jk} f(In(j)) \tag{10}$$

where θ_j and θ_k are the biases in the hidden layer and the output layer, respectively. The output value will be compared with the target by the mean absolute error as the error function

$$E_m = \frac{1}{2n} \sum_k \sqrt{(T_k - Y_k)^2} \tag{11}$$

Where the n is the number of training patterns, Y_k and T_k are the output value and the target value. The weight is adjusted according to the following expression:

$$w(t+1) = w(t) - n\partial E(t)/\partial w(t) \tag{12}$$

where t is the number of epochs and n is the learning rate.

The learning NN algorithms methods can be broadly divided into supervised, unsupervised and reinforced learning methods.

PREVIOUS STUDIES ON MACHINE LEARNING

The exponential growth of spam e-mails in recent years has resulted in the necessity for more accurate and efficient spam filtering. Machine learning (ML) is a very effective approach that has been successfully used in text classification. This approach is increasingly being applied to combat spam.

By allowing machines to classify e-mail into spam and non-spam messages, it relieves human intervention thus reducing the cost of monitoring spam.

Support vector machine (SVM) is one of the popular ML approaches being applied in anti-spam classification. Vapnik and his co-workers on 1999 initially applied this ML technique for spam classification. They tested it against three other techniques; Ripper, boosting decision tree and Rocchio. Both boosting trees and SVMs provided “acceptable” performance, with SVMs preferable given its lesser training requirements (Vapnik et al., 1999). The best result yield for SVM is obtained by using binary representation and a frequency-base for boosting.

The Naïve Bayesian (NB) approach was initially proposed by Sahami (1998) for automatic e-mail classification using decision theoretic framework and since the work, researchers have conducted many studies focusing on Naïve Bayes defeating spam. Androutsopoulos et al. (2000) investigated the effect of attribute-set size, training-corpus size, lemmatization, and stop-lists on the Naïve Bayesian filter’s performances. They concluded that after introducing cost-sensitive evaluation, additional safety nets are needed for the Naïve Bayesian anti-spam filter to be viable in practice. Graham (2002, 2003) later implemented a Bayesian filter that caught 99.5% of spam with 0.03% false positives.

Kun-Kan Li and et al (2002) classified spam using Simplified Support Vector Machine using pool-based active learning which involves selecting a training set of examples from a pool of unlabeled examples. Soonthornphisaj et al. (2002) investigated spam classification using a Centroid-Based approach in which the data items are represented using a vector space model, Naïve Bayesian and K-nearest Neighbor (kNN). Their result concluded that Centroid-Based classifier outperformed Naïve Bayesian and kNN.

Clark et al. (2003) classified spam using LINGER, a neural network-based system which uses a multi-layer perceptron. LINGER includes 2 feature selectors: Information gain (IG) and variance (V). Their results show that neural network-based filters achieve better accuracy in the training phase but has unstable portability across different corpora (Clark et al., 2003). Woitaszek et al. (2003) used simple SVM along with a personalized dictionary for model training. They subsequently implemented the classifier as an add-in for Microsoft Outlook XP providing sorting and grouping capabilities using Outlook’s interface to the typical desktop e-mail user.

Matsumoto et al. (2004), described the results of an empirical study on two spam detection methods: Support Vector Machines (SVMs) and Naive Bayesian Classifier (NBC). They used both term frequency (TF) and term frequency with inverse document frequency (TF-IDF) for features vector construction. Their results reflect that Naïve Bayesian has a consistent performance for all the data sets ranging.

Zhao and Zhang (2005) implemented a rough set based model to classify e-mails into three categories: Spam, non-spam and suspicious and compared it with Naïve Bayesian Classifier. The result shows that the Rough Set-based method has a better accuracy rate than that of the Naïve Bayesian. Chuan et al. (2005) proposed the use of LVQ-based neural network for spam e-mail classification. E-mails are classified into several subclasses for easy identification and Learning vector quantization (LVQ)-based NN. Their experiments shows LVQ-NN has better precision and recall rates compared to NN-BP and Naïve Bayesian in which Naïve Bayesian shows the lowest rates.

Wang et al. (2006) used the integration of two linear classifiers, Perceptron and Winnow. They concluded that Winnow produces slightly better results than Perceptron, however both classifiers performed very well and considerably outperformed the Naïve Bayesian classifier.

Ichimura et al. (2007) propose self organizing map (SOM) for spam classifications and automatically defined group (ADG) to extract correct judgment rules. They used 3007 e-mails classified as spam from SpamAssassin, SOM is used to classify these spam to obtain the visual distribution intuitively and the ADG extracted classification rules to judge spam correctly. Their experiment concluded that SOM improves the classification process and ADG tremendously reduces false negatives. Yang and Elfayoumy (2007) evaluated the effectiveness of feedforward backpropagation Neural Network and Bayesian classifiers for spam detection. Their result concluded that feedforward backpropagation NN provides relatively high accuracy compared to Bayesian classifier.

Lobato and Lobato (2008) used binary classification based on an extension of Bayes point machines. By using the Bayesian approach with inference expectation propagation (EP) they produced a result that outperforms SVM. Ye et al. (2008) proposed spam discrimination model based on SVM and the D-S theory. They used SVM with probability to sort out mail according to the features of mail headers and mail body textual content and D-S Theory to identify spam which improves the accuracy of the spam filter. Yu and Xu (2008) compared four ML algorithms; Naïve Bayes (NB), neural network (NN), support vector machine (SVM) and relevance vector machine (RVM). Their experimental results show that NN classifier is more sensitive to the training set size and unsuitable for using alone as spam rejection tool, SVM and RVM are superior to NB, and RVM is much faster testing time.

Wu (2009) used a hybrid method of rule-based processing and back-propagation neural network for spam filtering. A rule-based process is first employed to identify and digitize the spamming behaviors observed from the headers and syslogs of e-mails. Then they utilize the spamming behaviors as features for describing e-mails. This information is then used to train the BPNN. The system produced very low false positive and negative rates and with better results in comparison to content- based classification (Guzella, 2009).

Wang et al. (2009) developed and experimented anti-spam filtering system by combining Naïve Bayesian with

Table 4. Summaries of previous studies on ML Algorithms used and accuracy (English Language).

Reseacher (s)	Algorithm used	Accuracy (%)	False positive
Soonthornphisaj et al. (2002)	Centroid-Based approach	83	NA
Graham (2002, 2003)	Bayesian filter	99.5	0.03% FP
Woitaszek et al. (2003)	Simple support vector machine with personalized dictionary	95.26	6.80% FP
Zhao and Zhang (2005)	Rought Set Based	97.37	NA
Chuan et al. (2005)	LVQ-based Neural Network	98.97	NA
Wang et el. (2006)	Perceptron	98.89	NA
	Winnow	99.31	
Yang and Elfayoumy (2007)	Feed forward back propagation neural network	90.24	0.81% FP 0.84% FN
Lobato and Lobato (2008)	SVM and D-S Theory	98.35	NA
Sun et al. (2009)	LPP and LS-SVM	94	NA
Meizhen et al. (2009)	Behavior recognition based on fuzzy decision tree (FDT)	97	NA

distributed checksum clearinghouse (DCC) to avoid excessive false positives. This combination achieved very high recall, accuracy rates and exhibits excellent reliability and efficiency.

Yong et al. (2009) proposed anti-spam filtering based on the fuzzy clustering algorithm instead of classification algorithm that performs filtering without advance training processes. They calculated normalized cost (NC) by setting 11 values of threshold t, from 0.0 to 1.0 and λ = 1, λ = 9, λ = 999. The result suggests that to set a high value of threshold *t* achieves lower cost in the scenarios where over-blocking is severely punished.

Sun et al. (2009) proposed locality pursuit projection (LPP) and least square version of SVM (LS-SVM) for anti-spam whereby the LPP algorithm is used to extract features from e-mail and then classified as legitimate or spam mails by using the LS-SVM classifier. The result of their study shows much better performance compared to other classifiers.

Meizhen et al. (2009) proposed a model for spam behavior recognition based on fuzzy decision tree (FDT). This model can efficiently detect and analyze spammers' behavior patterns, and classify e-mails automatically. The system computed information gain to analyze and select behavior features of e-mails. They concluded that since absolutely clear attributes does not always exist in the real world, the attribute subordinating degree is more natural and reasonable to describe the characteristics of behavior. Fuzzy decision tree is more adaptive than Crisp decision tree.

The results show better accuracy rate with detection rate of more than 70% which indicates that the fuzzy decision tree model is a good prospect and efficient. Table 4 shows the summaries of ML Algorithms used and its accuracy.

There are many studies carried out on spam filtering that are effective and efficient on detecting and blocking spam e-mail, however these studies mainly performed based on English language-based (e-mail) spam. Methods (preprocessing and ML algorithms) used for English language spam detection may not produce higher performances given the nature of different human languages (Table 5).

PREPROCESSING

An e-mail is divided into a header section and a body section. The header section contains general information such as sender's information, recipient(s) information, subject and route information, where else the body contains the actual message. This information needs to be extracted before running a filter process by means of preprocessing. The purpose for preprocessing is to transform messages in mail into a uniform format that can be understood by the learning algorithm (Zhang et al., 2004).

The steps involved in preprocessing are as follows:

1. Feature extraction (Tokenization): Extracting features from e-mail; header or e-mail body into a vector space.
2. Feature selection: Dimensionality reduction; reduction of the features vector.
3. Stop word removal: Removal of non-informative words.
4. Noisy removal: Removal of obscure text or symbols from features
5. Features representation of features into appropriate format for the ML filtering.

Table 5. Other Languages and algorithms.

Author (s)	Language	Classifier algorithms used	Accuracy (%)
Ozgur et al. (2004)	Turkish	Artificial neural network and Bayesian	90
Dong et al. (2006)	Chinese	Bayesian spam filter based on cross N-gram	93
Pang Xiu-Li et al. (2007)	Chinese	Support vector machine based tri-gram language model and discount smoothing	98
Tuah et al. (2008)	Vietnamese	Vietnamese segmentation based on language classification and Bayesian	98.5
Na Songkhla and Piromsopa (2010)	Thai	Statistical rule-based	80.8
Qiu et al. (2010)	Chinese	Online linear classifier; Perceptron, Winnow and Naïve Bayesian	 97.56 97.33 94.2

TOKENIZATION

Tokenization is the process of reducing a message to its colloquial component (Zdziarski, 2005). It takes the message and breaks it up into a series of tokens (words). The words are obtained from the e-mail's message body although the header and subject fields also can be considered. These words/features are then added to a vector space to construct a features space for classification. The tokenization process will extract all the features from the message without regard of its importance. Tokenized features are highly vulnerable to content obscuring (Guzella and Carminhas, 2009) thus dimensional reduction, stemming and stop-word removal processes are required.

DIMENSIONALITY REDUCTION TECHNIQUES

The size of vectors containing the original features may be too large for a filter to handle. As such a dimensionality reduction technique is applied to the feature vector. There are many types of dimensionality reduction techniques mainly document frequency (DF), information gain, Chi-square, mutual Information, term strength, lemmatization and stop-word removal.

Document frequency

Document frequency refers to the number of documents in which a feature occurs. The weight of the features is measured in terms of frequency and the lower frequency, that is less than the predetermined threshold, is removed. Insignificant features that do not contribute to classification are ignored thus improving the efficiency of the classifier. The mathematical form of DF is as follows:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \qquad (13)$$

Information gain

Information gain is based on the feature's impact on decreasing entropy (Chen et al., 2008). IG measures the number of bits of information obtained for the category by knowing the presence or absence of a term in a document. Let $\{c_i\}_{i=1}^{m}$ denote the set of categories in the target space. IG of term t is defined as:

$$G(t) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + P_r(t)\sum_{i=1}^{m} P_r(c_i|t)\log P_r(c_i|t) + P_r(t)\sum_{i=1}^{m} P_r(c_i|t)\log P_r(c_i|t) \qquad (14)$$

Chi-square

Chi-Square is a statistical test that measures the occurrence of features against the expected number of the occurrences of those features (Yerazunis et al., 2005). In the Chi-square, the independent variables are the features and the dependent variables are the categories (that is legitimate and sp0am e-mail).

$$X^2(t,c) = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (15)$$

The above formula measures the term-goodness, a term t and a category c, where A is the number of times the t and c occurs together, B is the number of times the t occurs without c, C is the number of times c occurs without t, and D is the number of times neither c nor t occurs. The chi-square formula for category computation is as follows:

$$X^2_{avg}(t) = \sum_{i=1}^{m} P_r\,(c_i) X^2(t,c_i) \quad (16)$$

$$X^2_{max}(t) = \max_{i=1}\{X^2\,(t,c_i)\} \quad (17)$$

Mutual information

Mutual Information is a quantity that measures the mutual dependence of the two variables. If a feature x does not depend on a class c then it is removed from the vector space. For each feature attribute X with the category variable C, MI can be computed as follows:

$$MI(X;C) = \sum_{x\in\{0,1\},c\in\{spam,legit\}} P(X=x,C=c) * log\frac{P(X=x,C=c)}{P(X=x)*P(C=c)} \quad (18)$$

Mutual Information is accurate in predictions and an easier model to implement.

Stemming

Stemming is a process of reducing words to its basic form by stripping the plural from nouns (e.g. “apples” to “apple”), the suffixes from verbs (e.g. “measuring” to “measure”) or other affixes. Originally proposed by Porter on 1980, it defines stemming as a process for removing the commoner morphological and in-flexional endings from words in English. A set of rules is applied iteratively to transform words to their roots or stems. This approach reduces the number of features in the space vector and increases the learning speed and categorization phases for many classifiers. However, stemming may cause two different words to be stemmed as a same word.

Stop-word removal

Stop-word removal is the removal of common words that have high frequency but carries less meaning than the keywords. E-mail messages consists large number of non-informative words, such as articles (e.g. “a”, “an” and “the”), prepositions (e.g. “with” or “beside”) and conjunctions (e.g. “and”, “or” or “for”) and these words will increase the size of the vector space thus complicating the categorization process. A list of stop words is generated and is then compared against the space vector to eliminate words that are mapped to the list.

Noise removal

Obfuscated words in an e-mail represent noise. A deliberate action of misspelling, misplaced space or embedding special characters into a feature is referred to as obfuscation. For an instance, spammers obfuscated the word Viagra into “V1agra”, “V|iagra” or Free into “fr33”. Spammers employ this technique in an attempt to bypass the correct identification of these terms by spam filters (Guzella and Carminhas, 2009). Regular expression and statistical de-obfuscation techniques is used to contrast these misspelled terms.

REPRESENTATION

Features extracted from the e-mail are usually represented as a vector space model (VSM) or “bag of words”. The lexical features are represented in either binary or numeric. Vector space model represents message as vectors $\vec{x} = (x_1, x_2, \dots x_n)$, where $x_1, \dots, x_n$ are the values of attribute $X_1,\dots, X_n$. All attributes are binary: $X_i = 1$ if the corresponding feature/word is present in the message; otherwise $X_i = 0$. The numeric representation of the attributes where X_i is a number indicates the frequency of occurrence of the feature in the e-mail. For example if the word “Viagra” appears in the message then a binary value 1 will be assigned to the feature.

Another commonly used feature representation is character *n*-gram model which obtains sequences of characters and term frequency-inverse document frequency (tf-idf). n-gram is n-characters slice of a word. It also can be referred as any co-occurring set of characters in a word. n-gram encompasses bi-gram, tri-gram and qua-gram. tf-idf is a statistical measure used to calculate how significant a word is to a document in a feature corpus. Word frequency is established by term frequency (tf), number of times the word appears in the message yields the significance of the word to the document. The term frequency then is multiplied with inverse document frequency (*idf*) which measures the frequency of the word occurring in all messages (Robertson, 2004).

PERFORMANCE MEASUREMENT

Classifiers need to be evaluated based on the performance of information retrieval (recall, precision and derived measures) and decision theory (false positives and false negatives) (Guzella and Caminhas, 2009). Accuracy, spam precision and spam recall are the most important performance parameters. Recall indicates the number of correctly classified spam against spam that is misclassified as legitimate and the number of spam recognized as spam.

Precision represents the ratio between the numbers of correctly classified spam to the number of all messages marked as spam. Accuracy represents the ratio between the number of correctly classified spam and legitimate mails to the total e-mails used for testing that is all e-mails that are correctly classified by the classifier. These parameters can be measured using the following equations:

$$Accuracy\ (A) = \frac{TP+TN}{TP+TN+FP+FN} \tag{19}$$

$$Precision(p) = \frac{TP}{TP+FP} \tag{20}$$

$$Recall(r) = \frac{TP}{TP+FN} \tag{21}$$

$$F_i = \frac{2*r*p}{r+p} * 100\% \tag{22}$$

Spam e-mails that are classified as legitimate e-mail are referred to as false negatives (FNs) where else legitimate e-mail classified as spam is referred to as false positives (FPs). True positive (TP) means spam e-mails that correctly predicted as spam; True negative (TN) is the number of e-mail that is legitimate and is truly predicted as legitimate.

Based on this research, naïve Bayesian and neural network show promising and better techniques that can be applied to combat spam.

CONCLUSION

Spam is becoming one of the most annoying and malicious additions to Internet technology. Traditional spam filter software are unable to cope with vast volumes of spam that slip past anti-spam defenses. As spam problems escalate, effective and efficient tools are required to control them. Machine learning approaches have provided researchers with a better way to combat spam. Machine learning has been successfully applied in text classification. Since e-mail contains text, the ML approach can be seamlessly applied to classified spam. E-mail now can be classified with less human intervention thus making the control easier and more accurate. The effectiveness of a spam filter can be increased with preprocessing steps that are applied to the training and testing of features vectors.

Based on this research, naïve Bayesian and neural network show promising and better techniques that can be applied to combat spam. Researchers are planning to implement naïve Bayesian and neural network techniques to filter spam for Malay language e-mails.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers for their comments which helped to improve the presentation of the paper.

REFERENCES

Amayri O, Bouguil N (2009). Online Spam Filtering Using Support Vector Machines. IEEE., pp. 337- 340.

Androutsopoulos I, Koutsias J, Chandrinos KV, Paliouras G, Spyropoulos C (2000). An evaluation of naïve Bayesian anti-spam filtering. Proc. Of the workshop on machine learning in the new information age: 11th Europe conference on machine learning, pp. 9-17.

Chen C, Tian Y, Zhang C (2008). Spam Filtering with Several Novel Bayesian Classifiers. IEEE.

Chuan Z, Xianliang L, Mengshu H, Xu Z (2005). A LVQ-based neural network anti-spam e-mail approach. ACM SIGOPS Operating Syst. Rev., pp. 34-39.

Clark J, Koprinska I, Poon J (2003). A neural network based approach to automated e-mail classification. IEEE International conference on Web Intelligence, pp. 702-705.

Dalkilic G, Sipahi D, Ozcanhan MH (2009). A simple yet effective spam blocking method. Proceedings of the 2nd international conference on Security of information and networks, pp. 179-185.

Dong YS (2004). A comparison of several ensemble methods for text categorization. IEEE international Conference on Services Computing.

Dong J, Cao H, Liu P, Ren L (2006). Bayesian Chinese Spam Filter Based on Crossed N-gram. Intelligent Systems Design and Applications, 2006. ISDA '06. Sixth International Conference on, 3: 103-108.

Ferris Reseacrh (2009). Spam, Spammers, and Spam Control A White Paper by Ferris Research (March 2009). Retrieved 8th Feb 2010 http://apac.trendmicro.com/imperia/md/content/us/pdf/products/enterprise/interscanmessagingsecuritysuite/wp01_antispamferris_090311us.pdf.

Ferris Reseacrh (2010). Industry Statistics. Retrieved 8th Feb 2010 http://www.ferris.com/research-library/industry-statistics/.

Frost K, Udsen H (2006). Anti-spam regulation in Demark. Computer Law and Security, 22: 241-249.

Goyal RD (2007). Knowledge Based Neural Network For text classification. IEEE International Conference on Granular Computing, pp. 542-247.

Graham P (2002). A Plan for spam. Retrieved from http://www.paulgraham.com/spam.html.

Graham P (2003). Better Bayesian Filtering. Retrieved 10th March 2010 http://www.paulgraham.com/better.html.

Green T (2005). How URL Spam Filtering Beats Bayesian/Heuristics Hands Down. Retrieved 10th March 2010 http://www.spamstopshere.com/resources/documents/whitepaper.html?task=doc_download&gid=2.

Guzella TS, Caminhas WM (2009). A review of machine learning approaches to spam filtering. Expert system with Application, 36: 10206-10222.

Heron S (2009). Technologies for spam detection. Network Security Jan 2009, pp. 11-15.

Hideo A (2009). Study Report of an Anti-spam System with a 99% Block Rate (nov 2009). Retrieved 27th Feb 2010 from http://gabacho.reto.jp/en/anti-spam/anti-spam-system.html#1.

Ichimura T, Hara A, Kurosawa Y (2007). "A classification method for spam e-mail by Self-Organizing Map and automatically defined groups," Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on, pp. 2044-2049.

Ko M, Tiwari A, Mehnen J (2009). A review of soft computing applications in supply chain management. Applied Soft Computing, In Press, Corrected Proof, 15 September 2009, pp. 1-14.

Kun L, Kai L, Kuanwang H, Fengtian S (2002). Active Learning With Simplified Svms For Spam Categorization. Prdceedings of the First International Conference on Machine Laming and Cybernetic, pp. 1198-1202.

Lai CC (2007). An empirical study of three machine learning methods for spam filtering. Knowledge-Based System. Elsevier, 20: 249-254.

Lai GH, Chen CM, Laih CS, Chan T (2009). A collaborative anti-spam system. Expert Systems with Application. Elsevier, 36: 6645-6653.

Lazzari L, Mari M, Poggi A (2005). A collaborative and multi agent approach to e-mail filtering. IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'05), pp. 238-241.

Lobato DH, Lobato JM (2008). Bayes Machines for binary classification. Pattern Recognition Letters. Elsevier, 29: 1466-1473.

Matsumoto R, Zhang D, Lu M (2004). Some Empirical Results on Two Spam Detection Methods. IEEE, pp. 198-203.

Meizhen W, Zhitang L, Sheng Z (2009). A Method for Spam Behavior Recognition Based on Fuzzy Decision Tree. IEEE, Ninth International Conference on Computer and Information Technology, pp. 236-241.

MessageLabs intelligence (2005). Annual Security Report. (2005) Retrieved 8th Feb 2010 http://www.messagelabs.co.uk/intelligence.aspx.

MessageLabs intelligence (2009) Annual Security Report. (2009) Retrieved 8th Feb 2010 http://www.messagelabs.co.uk/intelligence.aspx.

Moustakas E, Ranganathan C, Duquenoy P (2010). Combating spam through legislation: a comparative analysis of us and European approaches. Retrieved 10th Feb http://www.ceas.cc/2005/papers/146.pdf.

Na Songkhla C, Piromsopa K (2010). Statistical Rules for Thai Spam Detection. Future Networks, 2010. ICFN '10. Second International Conference on Future Networks, 22-24: 238-242.

Oda T, White T (2003). Increasing the Accuracy of a Spam Defecting Aritical immune System. IEEE, pp. 390-396.

Ozgur L, Gungor T, Gurgen F (2004). Spam Mail Detection Using Artificial Neural Network and Bayesian Filter. Springer, pp. 505-510.

Pang X, Feng Y, Jiang W (2007). A Chinese Anti-Spam Filter Approach Based on Support Vector Machine. Management Science and Engineering, 2007. ICMSE 2007. International Conference on, 20-22: 97-102, Aug. 2007.

Porter MF (1980). An Algorithm for suffix Stripping. Available at : http://tartarus.org/~martin/PorterStemmer/def.txt.

Qiu Y, Xu Y, Wang B (2010). An Online Linear Chinese Spam E-mails Filtering System. E-Business and Information System Security (EBISS), 2010 2nd International Conference, pp.1-4: 22-23.

Rish I (2001). An empirical study of the naïve bayes classifier. Available at: http://www.cc.gatech.edu/~isbell/classes/reading/papers/Rish.pdf

Robertson S (2004). Understanding Inverse Document Frequency: On theoretical arguments for ID; J. Doc., 60(5): 503-520.

Sahami M, Dumais S, Heckerman D, Horvitz E (1998). A Bayesian Approach to Filtering Junk E-mail. In Learning for Text Categorization – Papers from the AAAI Workshop, pp. 55-62, Available at: ftp://ftp.research.microsoft.com/pub/ejh/junkfilter.pdf.

Schaub MY (2002). Unsolicited E-mail: Does Europe allow spam? The state of the art of the European legislation with regard to unsolicited commercial communications. Computer Law and Security Report. 18(2): 99-105.

Sebastiani F (2002). Machine Learning in Automated Text categorization. ACM Computing Surveys, 34(1): 1-47.

Soonthornphisaj N, Chaikulseriwat K, Tang-On P (2002). Anti-Spam Filtering: A Centroid-Based Classification Approach. IEEE, ICSP'02 Proceedings, pp. 1096-1099.

Spamhaus (2010). Definition of Spam. Retrieved 8th Feb 2010 http://www.spamhaus.org/definition.html.

Sun A, Lim EP, Ng WK (2002). Web Classication Using Support Vector Machine Proceedings of the 4th Int. Workshop on Web Information and Data Management (WIDM 2002).

Sun X, Zhang Q, Wang Z (2009). Using LPP and LS-SVM For Spam Filtering. ISECS International Colloquium on Computing, Communication, Control, and Management, pp. 451-454.

Tuah AN, Quang AT, Ngoc BT (2008). Vietnamese spam detection based on language classification. Communications and Electronics, 2008. ICCE 2008. Second International Conference on, 4(6): 74-79.

Vapnik VN, Druck H, Wu D (1999). Support Vector Machines for Spam Categorization. IEEE Transactions On Neural Networks, 10(5): 1048-1054.

Wang B, Jones GJF, Pan W (2006). Using online linear classifiers to filter spam e-mails. Pattern Analysis and Applications, 9: 339–351.

Wang H, Zhou R, Wang Y (2009). An anti spam filtering based on the Naïve Bayesian Classifier and Distributed Checksum Clearinghouse. 3rd International Symposium on Information Technology Application. IEEE computer society. pp. 128-131.

Woitaszek M, Shaaban M, Czernikowski R (2003). Identifying Junk Electronic Mail in Microsoft Outlook with a Support Vector Machine. IEEE Proceedings of the 2003 Symposium on Applications and the Internet (SAINT'03), pp. 166.

Wood P, Bleaken D, Nisbet M, Zhang J, Johnston N, Lee M, Lewis D (2010). MessageLabs intelligence: (2009) Annual Security Report. (2009) Retrieved 8th Feb 2010 http://www.messagelabs.co.uk/intelligence.aspx

Wosotowsky A, Winkler E (2009). Spam Report McAfee Labs Discovers and Discusses Key Spam Trends (2009). Retrieved 10th Feb 2010 http://www.mcafee.com/us/local_content/reports/7736rpt_spam_1209.pdf

Wu CH (2009). Behavior-based spam detection using a hybrid method of rule-based techniques and neural network. Expert Systems with Application. Elsevier, pp. 4321-4330.

Wu J, Deng T (2008). Research in Anti-Spam Method Based on Bayesian Filtering. IEEE, Pacific-Asia Workshop on Computational Intelligence and Industrial Application, pp. 887 – 891.

Xie M, Yin H, Wang H (2006). An Effictive defense against e-mail spam laundering. ACM. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.9653&rep=rep1&type=pdf.

Yang Y, Elfayoumy S (2007). Anti-Spam Filtering Using Neural Networks and Baysian Classifiers. Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation, pp. 272-278.

Ye M, Jiang QX, Mai FJ (2008). The Spam Filtering Technology Based on SVM and D-S Theory Workshop on Knowledge Discovery and Data Mining. IEEE, pp. 562-565.

Yerazunis SW, Chhabra S, Siefkes C, Assis F, Gunopulos D (2005). A unified model of Spam filtration. Mitsubishi Electric research laboratories. Retrieved 15th March 2010 http://www.merl.com.

Yong H, Guo C, Zhang X, Guo Z, Zhang J, He X (2009). An Intelligent Spam Filtering System Based on Fuzzy Clustering. IEEE, Sixth International Conference on Fuzzy Systems and Knowledge Discovery, pp. 515-519.

Youn S, McLeod D (2007). Efficient Spam E-mail Filtering using Adaptive Ontology. IEEE International Conference on Information Technolgy (ITNG'07), pp. 249-254.

Yu B, Xu Z (2008). A comparative study for content-based dynamic spam classification using four machine learning algorithms. Knowledge-Based Systems. ScienceDirect, 21: 355-362.

Yu J, Cheng F, Xiong H, Qu W, Chen XW (2008). A Bayesian approach to support vector machines for the binary classification. Neurocomputing. Volume 72, Issues 1-3, December 2008, pp. 177-185.

Zdziarski JA (2005). Tokenization: The Building Blocks of Spam. In

Pollock W, Zinkann E (Eds.), Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification San Francisco: No Starch Press, pp. 97-110.

Zhang L, Zhu J, Yao T (2004). An Evaluation of statistical spam filtering techniques. ACM Transaction on Asian Language Information Processing. 3(4): 243-269.

Zhao W, Zhang Z (2005). An E-mail Classification Model Based on Rough SetTheory. IEEE, pp. 403-408.