A Taxonomy of Email SPAM Filters

Hasan Alkahtani

Paul Gardner-Stephen

Robert Goodwin

A TAXONOMY OF EMAIL SPAM FILTERS

HASAN SHOJAA ALKAHTANI*, PAUL GARDNER-STEPHEN**, AND ROBERT GOODWIN**

* Computer Science Department, College of Computer Science and Information Technology, King Faisal University, P.O.
Box: 400 Al-Hassa 31982, Kingdom of Saudi Arabia
hsalkahtani@kfu.edu.sa

** School of Computer Science, Engineering and Mathematics, Faculty of Science and Engineering, Flinders University,
GPO Box 2100, Adelaide SA 5001, Australia
paul.gardner-stephen@flinders.edu.au , robert.goodwin@flinders.edu.au

Abstract:

SPAM email is well known problem for both corporate and personal users of email. Although SPAM has been well studied, both formally and informally, SPAM continues to be a significant problem. While the various anti-SPAM techniques have been described separately, the authors are not aware of any systematic presentation of them in the literature. We address this omission by presenting taxonomy of existing SPAM email countermeasures, and a brief description of the taxa we have proposed, and ascribe a number of existing SPAM filters to the various taxa.

Keywords*: *SPAM, Taxonomy, Email, Classification, Filters.

1. INTRODUCTION

Many researchers and technologists have worked for many years now to create systems with the goal of identifying and eliminating email SPAM before it reaches end users. Consequentially, a great diversity of SPAM filtering methods has been devised. To date, however, the authors are not aware of any systematic presentation of these various methods. This has the potential to lead to inconsistent terminology and a piece-meal approach to SPAM filtering, and difficulty in allowing researchers and users to identify the relationships between various SPAM filters. This may in turn result in suboptimal SPAM mitigation strategies where substantially similar SPAM filters are used in combination when the intended action may have been to combine SPAM filters to leverage their complementarity and orthogonality to improve the effectiveness of the resulting hybrid systems.

This paper seeks to address this void by presenting a brief taxonomy of major SPAM filtering methods in the hope that it will be an aid to participants in the fight against SPAM. While space limitations prevents this taxonomy from being exhaustive, it is hoped to be representative of existing SPAM filters and illustrative of the major approaches to filtering SPAM.

2. A TAXONOMY OF EMAIL SPAM FILTERS

Our taxonomy is depicted in Figure 1, with each taxon described in the text below.

2.1. REPUTATION BASED FILTERS

This major class of SPAM filter relies on information outside of the content of the individual email messages. These filters make assessments about the reputation of one or more of the participants (sender, recipient and intermediaries) in the email transaction. The various methods differ in the subject of the reputation calculation, e.g. IP address, sender domain and sender address, and also in the nature of the reputation calculation. We divide reputation based filters into three major techniques which are: (a) origin based; (b) social based, and; (c) traffic analyzing.

2.1.1 ORIGIN BASED TECHNIQUES

Origin based techniques classify SPAM based on network information, such as the source IP and email addresses. Such analyses have the advantage that they can be performed before an email is received by recipients, potentially saving network and computational resources [12]. Numerous origin based techniques exist, including: (1) Black lists; (2) White lists; (3) Challenge-Response Systems, and; (4) origin diversity analysis [8].

Black Lists: Black lists include the Realtime Blackhole Lists (RBL) and Domain Name System Blacklists (DNSBL) [8]. These databases list IP addresses of suspected spammers or known spammers. With these black lists SPAM can be blocked at the SMTP connection phase.

However, while black lists are reasonably effective and efficient, they have disadvantages. First, black lists are maintained by an entity distinct from the user, introducing an external dependency into any SPAM filter that relies on them. Second, the effectiveness of black lists depends on the timeliness and methods of those who manage them. Finally, because most black lists are usually queried via DNS, this can result in substantial DNS traffic and consequent delays in SPAM

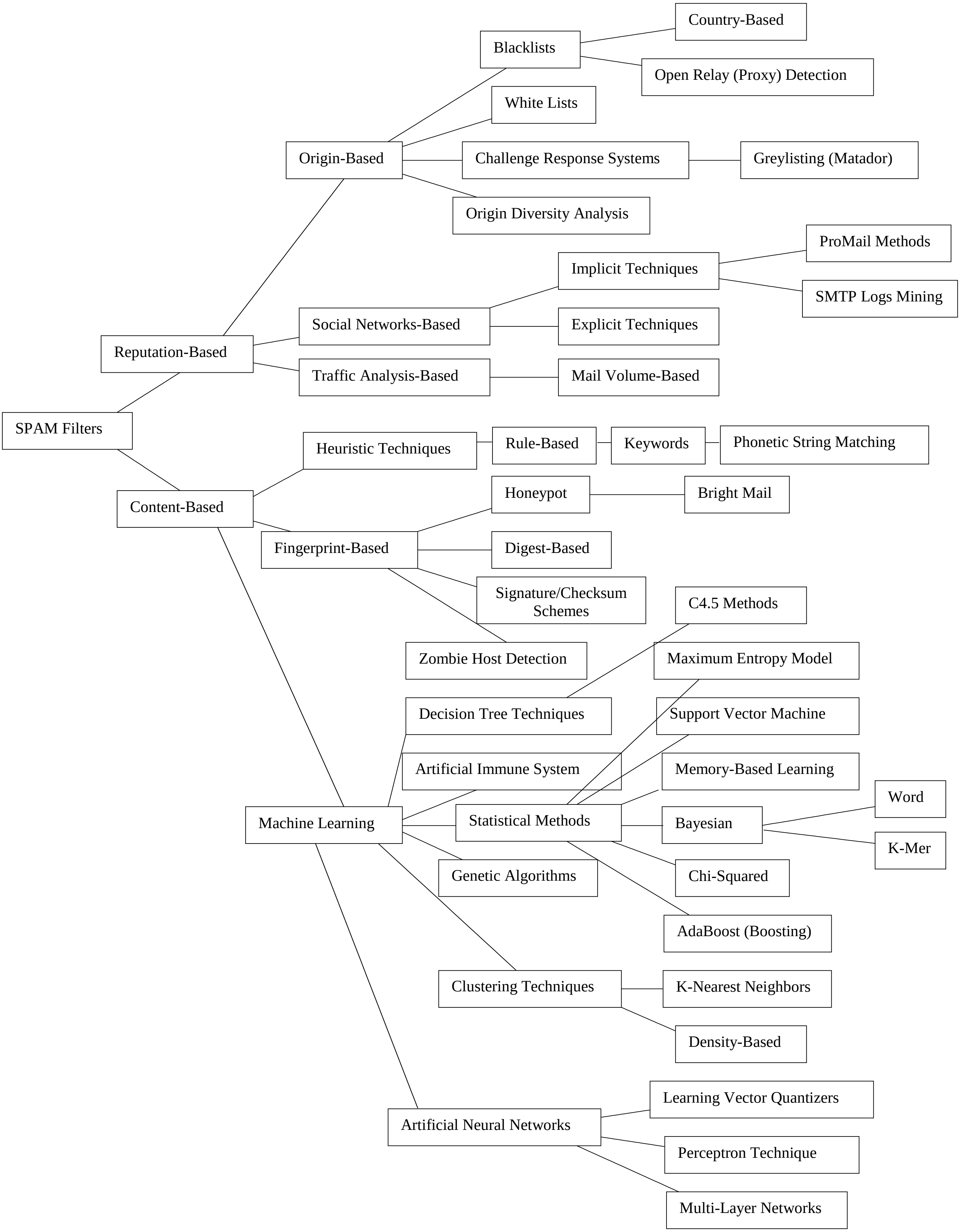

Figure 1: A Taxonomy of Email SPAM Filters.

processing, especially for mail servers that reference more than one blacklist [8].

Many methods are used to produce and maintain black lists, including open relay detection and country-based techniques. First, an open relay or proxy is a SMTP server that indiscriminately relays all email it is presented with, and without validating the headers of those messages. Such servers allow spammers to not only hide their origin, but also to add falsified headers to misdirect anyone seeking to identify their source [2]. For this reason several black lists routinely list all open proxies that they identify [12]. Second, some countries are considered particularly notorious sources of SPAM. Some black lists thus indicate that mail should be rejected from any SMTP server in that country, with the resulting false-positive problem [17].

Indeed, false positives are a problem for many SPAM countermeasures, and thus it is common to combine the opinion of multiple filtering methods before making a final classification.

White Lists: White lists enable users to create a list of trusted addresses which they nominate for exclusion from SPAM filtering. Email originating from all others addresses will be filtered as normal [21]. White lists reduce the cost of filtering SPAM, because some email messages are allowed to bypass the SPAM filter process [8]. However, white lists require user interaction. Also, many white list implementations rank white listed email above all other email in the inbox. Thus white lists can make it difficult to identify email that is not white listed, but is nonetheless legitimate, due to its reduced visibility, especially when a large proportion of SPAM is present [14].

Challenge-Response Systems (CRS): Whereas white lists place the burden on the receiver for determining trustworthy senders, challenge-response systems transfer the burden of authentication to the sender. The sender will receive an automated reply or challenge which require some action to prove that they are real users. For example, the system might send an image which contains a picture of animals for the sender. The sender is requested to count the number of animals in the picture. Such tasks are chosen to be trivial for a real person, but too difficult for a computer to perform quickly enough to facilitate effective spamming [20].

Thus, one advantage of challenge-response systems is that they protect against automated SPAM sending programs. A second advantage of challenge-response systems is that they require that the SPAM originate from a functional mailbox that is monitored by the spammer. If this were so, it would provide an easy means of identifying and filtering SPAM from that sender. Finally, challenge-response filters can be used to populate white lists [8].

A critical disadvantage of challenge-response filters is that if both sending and receiving mail servers implement them, dead lock will result as both servers will wait for the other to respond to their challenges. One way to reduce, but not eliminate, this dead-lock problem is to use challenge response systems in a grey listing filter. In this way the challenge is only send for email that is considered possibly SPAM.

Origin Diversity Analysis: Origin diversity analysis method is a hybrid origin and content based technique that focuses on the behavior of SPAM emails instead of the content of SPAM messages. This method clusters similar messages, and then considers the claimed origins of the messages. If there are many putative origins, then it is assumed that most of them must be fraudulent, and hence likely to be SPAM [13]. The claimed advantage of this scheme is that it relies only on a distinguishing behavior of SPAM in that it arrives in quantity from many apparent locations, since any deviation from that combination would reduce the effectiveness of the SPAM. Thus it has the potential to be robust in the face of the continuing evolution of SPAM.

2.1.2 SOCIAL FILTERS

There are many social network based proposed to combat SPAM. These methods all aim to assign to each message a probability of it being SPAM, based on the past history of the participants. Social filters are classified into: (a) implicit techniques, and; (b) explicit techniques.

Implicit Techniques: Social filters which were proposed by [5] are used to analyze fields of emails headers like ’To’ , ’Cc’ and ’Bcc’ to build a graph of social relations of users and classify new emails based on this graph [4].

ProMail and related methods construct a social network graph of email passing through an SMTP server, often by mining log files. Typically, nodes in the graph represent email accounts, while edges represent email transactions. These graphs are used to make decisions about whether a message is likely to be from a source in the recipients social network, and hence more likely to be legitimate [16, 26].

Explicit Techniques: In contrast to the implicit methods, there exist methods that explicitly build the social network through user interaction and may also utilize user-supplied or automatically computed reputation ratings [14]. These methods are naturally complementary with white listing and challenge response systems.

2.1.3 TRAFFIC ANALYSIS

Anomalies and patterns in the network traffic stream can be detected by mining the log files of an SMTP server. Although other analyses are possible, one common analysis that is used to detect SPAM is to identify when a host or network issues an abnormally large amount of email [12]. However, this technique results in a very high false acceptance rate [12].

2.2. CONTENT BASED FILTERS

In contrast to reputation based filters, content based filters detect SPAM by examining the content of email messages, irrespective of the origin.

Common Traits of these Techniques Include: (1) They require the body of a message before they can classify messages as SPAM or HAM, and thus incur the use of more network bandwidth compared with reputation based filters, and; (2) They are immune to the originating location of message, unlike origin-based techniques.

There exist several families of content based filtering techniques, including: (a) Heuristics; (b) Machine Learning, and; (c) Finger Printing.

2.2.1 HEURISTIC FILTERS

In these filters, email can be classified as SPAM by searching for patterns that have are commonly identified in SPAM. Patterns can be specific words, phrases, malformed message headers, exclamation marks and capital letters [8]. Perhaps the most common types of heuristic filters are the rule-based filters.

Rule Based Filters: Rule based filters were very common and popular until 2002 [8]. The classification of SPAM emails relied on user specified rules which characterize known unwanted emails. Rule based filters depend on the occurrence of critical words to classify SPAM. In addition, rule based filters do not only analyze the content of email, but also analyze email header which contain list of recipients, IP address's source and subject [11].

However, there is a problem with these filters which is word obfuscation. For example, a rule-based filter might have a rule to match the word 'Free', but that rule would not necessarily match the strings 'f*r*e*e' or 'bonus' [8].

One partial solution to word obfuscation is phonetic string matching, which provides a more robust pattern matching based on its phonetic transcription. This technique seeks to address the problems experienced by keywords-based filters when faced with word obfuscation [11].

2.2.2 MACHINE LEARNING FILTERS

Machine learning techniques aim to avoid the human labor required to maintain rule based filters by automatically deriving a HAM/SPAM classifier. By definition, these techniques need to be fed pre-classified training data, although once primed many can provide their own forward training to attempt to keep abreast of the evolution of SPAM. Several categories of machine learning SPAM filters are: statistical filters, genetic algorithms, artificial immune systems, artificial neural networks, clustering techniques and decision trees.

Statistical Filters: Statistical filters rely on a corpus of SPAM emails and legitimate emails to conclude features which can be used to classify incoming emails. If the statistical properties are closer to corpus of SPAM emails, the email is classified as a SPAM. Otherwise, email is classified as a legitimate if the statistical properties are closer to legitimate emails corpus. As with rule-based filters, many statistical filters also consider the header portion of messages [11]. A selection of statistical SPAM filters is: Bayesian, chi-squared, support vector machines (SVM), Boosting, maximum entropy models and memory-based learning techniques.

Bayesian SPAM filters consider the historical probability of each word in the message occurring in either SPAM or non-SPAM (HAM) messages [3, 22]. They calculate the probability that email is SPAM or non-SPAM by combine the individual SPAM/HAM probability of each words or k-mers [24] inside the messages to produce a final probability estimate that an email is SPAM or HAM (non-SPAM) [8].

The percentage of false positive generated by Bayesian filters are low, and they are self-adapting to stop new SPAM by receiving ongoing training form the user [12]. While extremely effective for a time, more recently Bayesian filters have become less effective due to the common practice of including random blocks of text into SPAM messages to reduce the accuracy of this detection technique.

[20] Have proposed a technique that uses the chi-based authorship identification technique to the SPAM identification problem, by applying the Chi by degrees of freedom test [4].

SVMs are a supervised learning method [27], used in text classification [18], that have more recently been applied to the SPAM identification problem [10].

Boosting is a learning algorithm which is based on the idea of combination of many weak hypotheses, for example as in the AdaBoost system [27]. A learner is trained in each stage of the classification procedure, and the output of each stage uses to reweight the data for the future stages [4]. Boosting algorithms with confidence-rated predictions have been proposed as being well suited to the SPAM filtering problem, and that they can outperform both Bayesian and decision tree methods [6].

Maximum entropy models are another machine learning technique from natural language processing that has also been applied to SPAM filtering [18, 27].

Memory based learning are "Non-Parametric Inductive Learning Paradigm that stores training instances in a memory structure on which predictions of new instances are based" [27] and have also been applied to the problem of SPAM [23].

Genetic Algorithms: Genetic SPAM detection algorithms use, feature detectors, that are often evolved over time, are used to score emails. The classification of emails as SPAM or non-SPAM is based on some integration of one or more such feature scores [12, 15].

Artificial Immune System: Artificial immune systems are machine learning methods used to fight SPAM and viruses of computers using methods that are in some way based on the immune system of biological organisms. The classification of emails to SPAM or HAM in this technique can be based, for example, on artificial lymphocytes created from a gene database, where the genes represent mini languages to include keywords which are checked in SPAM [18].

Artificial Neural Networks: ANNs are a common classification technique in artificial intelligence applications. ANNs represent networks of virtual neuron cells and are trained to perform some task. For SPAM detection, ANNs will typically classify incoming emails based on common features in emails [12]. There exist many types of ANNs [10], including: perceptrons, multi-layer networks, and learning vector quantizers (LVQ).

Perceptrons are generated by trying to find a linear function for some feature vector [18] which ideally produce distinct ranges of output values when given SPAM or HAM.

A multi-layer neural net is "a network of connected perceptrons which from a network with successive layers" [18], as such they are potentially more powerful than perceptrons.

Learning vector quantizers cultivate a set of neurons selecting the best neuron for each classification task and preens those neurons to increase their accuracy. LVQs are well suited to text classification tasks, and have been applied to the SPAM classification problem with superior results to both Bayesian and various other forms of artificial neural networks [7].

Clustering Techniques: Two examples of clustering techniques which have been applied to SPAM classification are K-nearest neighbors (KNN) and density-based clustering.

K-nearest neighbors (KNN) clustering indexes and converts emails to a high-dimensional vector and then measures the distance between the vectors of each email. Clusters are formed of neighboring, i.e., relatively close vectors. Once clusters have been formed SPAM classification need only be performed for a sub-set of any cluster population, as the result can then be inferred to apply to the other members of the cluster [10].

Density based clustering is another form of document clustering that has also been applied to SPAM classification [25]. A claimed advantage is the ability to process hashed versions of messages, thus preserving user privacy. These methods depend on having sensitive comparators. These comparators are usually either fast or sensitive. The challenge is finding comparators that are both sufficiently fast and sufficiently sensitive.

Decision Tree Technique: Decision trees are a classification technique commonly used in data mining, where the interior nodes of the tree represent observations, and leaf nodes decisions or conclusions. One decision tree technique that has been applied to SPAM classification is C4.5 [10].

2.2.3 FINGER PRINTING FILTERS

Finger printing methods make use of a list of "Finger Prints" of known types of SPAM, by computing and comparing the finger print of any incident email. Various schemes for generating finger prints are possible, e.g., via an exact or approximate digest or hashing algorithm [9, 12]. In any case, lists of the resulting finger prints are usually propagated to mail servers and any message, they receive that has a matching finger print is assumed to be SPAM [1]. Particular challenges in finger print based SPAM detection is making them reliable in the face of polymorphic SPAM, and ensuring that a finger print does not disclose or prove any content of a given message.

Honey Pots: A common method of collecting known SPAM messages for a finger printing system is via a honey pot, which is a machine or system that exists solely to collect SPAM [2]. Honey pots are also of value to researchers by identifying new species of SPAM as they emerge, as well as analyzing email harvesting activity, and detecting emails relays. BrightMail is an example of a honey pot.

BrightMail filters email addresses before placing them in the POP mailbox. It allows for spammers to detect email addresses left on web pages, news groups or subscription to mailing lists to send SPAM email to these addresses. In addition, BrightMail places spammers' email addresses in its blacklists databases. As a result, emails sent from these addresses are blocked.

Zombie-Based Approach: Spammers can send their emails by SPAM-bots or zombie machines [16]. Many zombie machines often use nonstandard optimizations to the SMTP protocol which can be detected by the receiving SMTP server. Thus it is possible to classify some SPAM based on the content of the SMTP session [19].

3. CONCLUSION

In this paper, we have presented a wide range of the techniques that have been used or proposed for use to fight SPAM, and attempted to indicate which SPAM filters use which techniques. We have sought to arrange these techniques in an orderly and informative manner, in the hope that the result will prove helpful in the continuing fight against SPAM, by allowing intelligent selection of SPAM filters by practitioners, and more consistent and informed treatment of SPAM filters in the academic literature compared with the previous situation.

The taxonomy presented here is clearly preliminary in nature, and non-exhaustive. A clear task for the future is to expand it with information about additional SPAM filters and techniques, and to address any refinements that become apparent during that process.

REFERENCES:

[1] Allman E., "Spam, Spam, Spam, Spam, Spam, the FTC, and Spam", *Queue*, vol. 1, no. 6, pp. 62-69, 2003.

[2] Andreolini M., Bulgarelli A. et al., "Honey Spam: Honey Pots Fighting Spam at the Source", *Proceedings of the Steps to Reducing Unwanted Traffic on the Internet on Steps to Reducing Unwanted Traffic on the Internet Workshop*, Cambridge, MA, pp. 1-13, 2005.

[3] Androutsopoulos I., Koutsias J. et al., "An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-Mail Messages", *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athens, Greece, pp. 160-167, 2000.

[4] Blanzieri E. and Bryl A., "A Survey of Learning-Based Techniques of E-Mail Spam Filtering", *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63-92, 2008.

[5] Boykin O. and Roychowdhury V., "Personal Email Networks: An Effective Anti-Spam Tool", *Condensed Matter cond-mat/0402143*, pp. 1-10, 2004.

[6] Carreras X. and Marquez L., "Boosting Trees for Anti-Spam E-Mail Filtering", *Proceedings of RANLP, 4th International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, BG, pp. 1-7, 2001.

[7] Chuan Z., Xianliang L. et al., "A LVQ-Based Neural Network Anti-Spam E-Mail Approach", *SIGOPS Oper. Syst. Rev.*, vol. 39, no. 1, pp. 34-39, 2005.

[8] Cook D., Hartnett J. et al., "Catching Spam Before it Arrives: Domain Specific Dynamic Blacklists", *Proceedings of the 2006 Australasian workshops on Grid computing and e-research - Volume 54*, Hobart, Tasmania, Australia, pp. 193-202, 2006.

[9] Damiani E., Vimercati S. D. C. d. et al., "An Open Digest-Based Technique for Spam Detection", San Francisco, CA, USA, pp. 1-6, 2004.

[10] El-Halees A., "Filtering Spam E-Mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques", *The International Arab Journal of Information Technology*, vol. 6, no. 1, pp. 52-59, 2009.

[11] Freschi V., Seraghiti A. et al., "Filtering Obfuscated Email Spam by Means of Phonetic String Matching", *ECIR*, pp. 505-509, 2006.

[12] Garcia F. D., Hoepman J.-H. et al., "SPAM FILTER ANALYSIS", *SEC*, pp. 395-410, 2004.

[13] Gardner-Stephen P., "A Biologically Inspired Method of SPAM Detection", *Database and Expert Systems Application, 2009. DEXA '09. 20th International Workshop on*, pp. 53-56, 2009.

[14] Golbeck J. and Hendler J., "Reputation Network Analysis for E-Mail Filtering", *CEAS*, pp. 1-8, 2004.

[15] GOWEDER A. M., RASHED T. E. , ALI S. et al., "An Anti-Spam System Using Artificial Neural Networks and Genetic Algorithms", *Proceedings of the 2008 International Arab Conference on Information Technology*, pp. 1-8, 2008.

[16] Hayati P. and Potdar V., "Evaluation of Spam Detection and Prevention Frameworks for E-Mail and Image Spam: A State of Art", *Proceedings of the 10th International Conference on Information Integration and Web-based Applications \& Services*, Linz, Austria, pp. 520-527, 2008.

[17] Heron S., "Technologies for Spam Detection", *Network Security*, 2009, 1, pp. 11-15, 2009.

[18] Khorsi A., "An Overview of Content-Based Spam Filtering Techniques", *Informatica (Slovenia)*, pp. 269-277, 2007.

[19] Lieven P., Scheuermann B. et al., "Filtering Spam Email Based on Retry Patterns", *IEEE International Conference on Communications*, pp. 1515-1520, 2007.

[20] O'Brien C. and Vogel C., "Spam Filters: Bayes vs. Chi-Squared; Letters vs. Words", *Proceedings of the 1st international symposium on Information and communication technologies*, Dublin, Ireland, pp. 291-296, 2003.

[21] Pfleeger S. L. and Bloom G., "Canning Spam: Proposed Solutions to Unwanted E-Mail", *IEEE Security and Privacy*, vol. 3, no. 2, pp. 40-47, 2005.

[22] Sahami M., Dumais S. Et al., "A Bayesian Approach to Filtering Junk E-Mail", *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, pp. 1-8, 1998.

[23] Sakkis G., Androutsopoulos I. et al., "A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists", *Information Retrieval*, vol. 6, no. 1, pp. 49-73, 2003.

[24] Sculley D., Wachman G. et al., "Spam Filtering Using Inexact String Matching in Explicit Feature Space with On-Line Linear Classifiers", *Text REtrieval Conference*, pp. 1, 2006.

[25] YOSHIDA K., ADACHI F. et al., "Density-Based Spam Detector", *Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, pp. 486-493, 2004.

[26] Lam H.-Y. and Yeung, D.-Y., "A Learning Approach to Spam Detection based on Social Networks", *Conference on Email and Anti-Spam, CEAS 2007*, pp. 1-9, 2007.

[27] Zhang L., Zhu J. et al., "An Evaluation of Statistical Spam Filtering Techniques", vol. 3, no. 4, pp. 243-269, 2004.