Approaches for Web Spam Detection

Kanchan Hans
Amity Institute of Information Technology
Amity University
Noida, India

Laxmi Ahuja
Amity Institute of Information Technology
Amity University
Noida, India

S.K. Muttoo
Department of Computer Science
University of Delhi
Delhi, India

ABSTRACT
Spam is a major threat to web security. The web of trust is being abused by the spammers through their ever evolving new tactics for their personal gains. In fact, there is a long chain of spammers who are running huge business campaigns under the web. Spam causes underutilization of search engine resources and creates dissatisfaction among web community. Web Security being a prime challenge for search engines has motivated the researchers in academia and industry to devise new techniques for web spam detection. In this paper we present a comprehensive survey of techniques for detection of web spam and discuss their applicability and performance in various scenarios where they outperformed the others. We have categorized web spam detection with the primary focus on the approaches used for spam detection. The paper also gives the possible directions for future work.

General Terms
Web Security, Search Engine, Anti-spamming, spam detection Quality Search

Keywords
Anti-Spam, web security, spam detection, approaches, search engines

1. INTRODUCTION
Adversarial Information retrieval has been an emerging issue of research for both academia and industry. World Wide Web is the huge repository of information satisfying needs of billions of users but at the same time there are adversaries also known as spammers who alter this information for their personal gains. They are the major threat to web security. The victims are the users who, on making query to web search engine are presented with unexpected pages which often are loaded with the malicious content. Such techniques employed by spammers are referred to as spamdexing [54]. Search engines also suffer as spam damages the reputation of search engines and increases cost in terms of storing, indexing and crawling spammed web pages. In order to give accurate results, search engines are actively handling this issue. Many researchers across the globe are working to mitigate spam, but at the same time the spammers also devise new tactics to evade the efforts made by SEO's and researchers.

Recent studies indicate that the amount of web spam is dramatically increasing. According to Symantec's Internet Security Threat Report [55], 2014, web based attacks have increased by 23%. Web Sense 2013 Threat Report [58] states that the number of malicious web links has grown by 600% in the last year which is an alarming situation. A Report produced by Mc. Café [33] shows that browser based attacks are leading all types of attacks. Also year 2013 has seen a rise in number of malicious URLs and domains by 22%. Given this scenario, this paper presents a comprehensive survey of various approaches that can be applied in arena of spam detection. There are some existing surveys [26][50] done in this area, but our work is different from them as it is more comprehensive and has focus on the approach used for spam detection. Also, we have discussed the applicability of these approaches and the situations in which they have outperformed the others. This survey will be help future researchers to have insight into the seriousness of the problem and will help them in framing the future directions in fighting spam.

The paper is organized as follows: the following section outlines the classification of web spam giving overview of spamming techniques. The next section explores various approaches developed for web spam detection. After that performance of various approaches is discussed and their applicability issues in different scenarios. In the end, we conclude the paper and discuss possible directions for future work.

2. CLASSIFICATION OF WEB SPAM
Spammers have been successful in devising new sophisticated techniques to spread spam. The main underlying motives are revenue generation, higher search engine ranking, promoting products & services, stealing information, and phishing. According to Gyongyi and Garcia-Molina [27], spamming techniques can be classified into two major categories: boosting techniques and hiding techniques.

2.1 Boosting Techniques
Boosting Techniques refer to all such techniques that are used by spammers to boost the rank of the page so that their websites can come in top results of search engine. It primarily includes content spam and link spam.

2.1.1 Content Spam
It refers to altering the textual content of the page by using a number of tricks. Traditionally, search engines used TF-IDF based algorithms of information retrieval that rank web pages on the basis of page content. Two quantities namely Term Frequency(TF) and the relation of a document d and a term t can be characterized by two quantities i.e. the number of times term t appears in document d and the other is Inverse Document Frequency (IDF), the ratio of the total number of documents to the number of documents that contain term t. Spammers can try to boost TF of terms. Spammers smartly analysed the weaknesses of these models and exploited them for creation of spam. Various tricks used by them as illustrated by Gyongi [27] are repetition, dumping, weaving, frame stitching and many more.

2.1.2 Link Spam
Link spam is the manipulation of the link structure or anchor text among pages to get a higher rank. Spammers generally

misuse the link-based ranking algorithms to achieve higher ranking for their spammed website. Spammers deceive ranking algorithms by creating densely connected set of pages. According to Zhang, link farm refers to "manipulation of the link structure by a group of users with the intent of improving the rating of one or more users in the group".

2.2 Page Hiding Techniques

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as distinguishing source code text. If Times Roman is not available, try the font named Computer Modern Roman. On a Macintosh, use the font named Times. Right margins should be justified, not ragged. Page Hiding Spam refers to methods

that by themselves do not influence the search engine's ranking algorithms, but that are used to hide the adopted boosting techniques from the eyes of human web users. In other words, it refers to methods intended to deceive web browsers and search engine experts by hiding web page or a part of the page which is not detected using visual inspection. It primarily includes cloaking and redirection.

2.2.1 Cloaking

Cloaking is a technique by which a Web server provides to the crawler of a search engine a page that is different from the one shown to regular users. It can be used legitimately to provide a better-suited page for the index of a search engine, for instance by providing content without ads, navigational aids, and other user interface elements. It can also be exploited to show users content that is unrelated to the content indexed.

2.2.1 *Redirection*

Redirection is a technique in which, the spam server automatically redirects the Web browser to another URL as soon as the page loads. This way the search engine still indexes the page, but the user never sees it. Pages with redirection are in essence intermediates (proxies or doorways) for the ultimate targets, which spammers try to serve to users through search engines. This is usually accomplished by using a scripting language such as JavaScript to redirect the user to a spam site. Most search engines do not interpret all the scripts due to the high computational cost of doing so for every page.

3. SPAM DETECTION APPROACHES

Detecting spam has always been a major challenge for the web community and interest area of researchers from both academia and industry. Lots of efforts have been done to combat with spam. Various approaches followed to detect this spam content are detailed as follows:

3.1 Machine Learning Approach

This approach requires designing the programs that learn from experience and try to detect patterns from data and perform classification. Machine Learning approaches are broadly classified into supervised and unsupervised learning. The primary difference between the two is that supervised learning algorithms require an initial training set for assistance in classification whereas it is not required in non- supervised algorithms. Various techniques employed under Machine Learning approach are Bayesian classification, neural-networks, Markov-based models, and pattern discovery. These machine learning techniques have been used to fight against content spam.

Ntoulas et al. [40] used machine learning methodology of detecting spam. They used C4.5 decision tree classifier [43] and used features based on page's content such as number of words in the page title, number of words in a page, average length of words, amount of anchor text, compressibility, fraction of visible content etc. In [21], authors used the ensemble classifier along LogitBoost [25] and Random Forest [12] to improve accuracy significantly. They compiled a minimal feature set that can be computed very quickly to allow intercepting spam at crawl time only. Amitay et al. [9] proposed using categorization algorithms to detect a website's functionality. Although their work was not aimed at detecting web spam, they identified clusters, each of which appeared to be a spam ring. In [13], Becchetti et al used automatic classifier for link spam detection. In another work, Becchetti et al. [14] designed classifier for link farms detection. They used some new features like Trust Rank value, Truncated PageRank value and estimated supporters in their classifier.

Silva et al. [47][48][49] has used different neural-based algorithms in detecting spam content on web. They also evaluated the performance of classifiers by modifying the feature vectors. Also, Najork has used machine learning for detecting cloaked pages [39]. He proposed an idea of detecting cloaked pages from users' browsers by installing a toolbar and letting the toolbar send the signature of user visited pages to search engines. A component called Search User Redirection Finder (SURF) [32] is designed as a browser component that extracts a number of features from browsing sessions, and based on this information identifies the malicious redirections. Their strategy is that during the session, SURF collects the information about browser events to track page (and frame) loads and redirections, network information to model the redirection chain and search result information for measuring the poisoning chances of the landing page. From this information, a number of statistical features are extracted, which are then fed to a classifier trained to identify instances of search poisoning. Kurt et. al [29] proposed a real time system called "Monarch" that determines whether the URLs direct to spam. This system crawl URLs as they are submitted to web services. It is based on features drawn from lexical properties of URLs, hosting infrastructure and page content and page behavior such as JavaScript events, plug-in usage and page's redirection behavior. They have used binary classifier to employ the training algorithm. Support vector machines (SVM) algorithm of machine learning too is used extensively [5], [6-8][57] in detecting spam patterns, classifying and other learning tasks [17][19].

3.2 Graph Based Approach

This approach considers the Web as a directed graph the set of Web pages form the vertices and the links between web pages act as edges. Web forms a bow-tie structure and is divided into five components based upon the properties of links. Properties of graph have been used in detection of spam. The methods for link farm detection search for anomalous patterns within the interconnection graph of the Web.

Wu and Davison [60] discovered link farms by first finding a candidate set of pages whose in-links and out-links have a sufficient number of domains in common. This list of candidates is then expanded by finding pages with sufficient links to confirm the cases of spam. Yu et al. [61] used random routes and his idea was that in a random walk, every time the

random walker arrives at a node he chooses at random which out-link to follow. Abernethy [1] proposed a novel approach for Web spam classification using graph-regularized classifiers. Akoglu [4] has devised methods for abnormality detection in graph data. Castillo et.al [16] used topology of web graph considering the link dependencies among the web pages to design a spam detection strategy. Recently Cohen and Kou [18] described a meta-learning scheme called stacked graphical learning. They used a base learning scheme C to derive initial predictions for all the objects in the dataset. Then, a set of extra features are generated for each object, by combining the predictions for the related objects in the graph using an aggregate function. Finally, this extra feature is added to the input of C, and algorithm is executed again to get new predictions for the data.

3.3 Trust or Badness Based Approach

In this approach, some initial known (labeled) pages are taken as seed set. The system is presented with a confirmed set of trustworthy and untrustworthy pages as inputs that are further used to compute the labels of other nodes on the basis of propagation rules. Such method uses the scoring mechanism where each node is assigned some trust or distrust score, which is propagated to next node. The basic idea behind the trust propagation is "the friend of a friend is my friend" and behind the distrust propagation is "the enemy of my enemy is my friend".

Many researchers have proposed different models of trust. Page Rank [51] assigns scores to the pages on the basis of information about in-degree of links. The underlying idea is that the popularity of a page is related to the in-degree of links i.e. a page is important if it has many other pages pointing towards it. Another algorithm named BadRank described in this paper works on the lines opposite to TrustRank. In this algorithm they considered badness to be propagated to the reachable pages. Initially a bad page set is selected and each page within the bad pages set is assigned a badness value. Becchetti developed a revised PageRank– Truncated PageRank algorithm to combat link based spam. The basic assumption is that for link farm spam pages, they may have lots of supporters within a few steps in the web graph, but little or no support at higher distances. Gy¨ongyi et al. proposed an algorithm, TrustRank, to combat link spam [62], assumes that good pages usually point to good pages and seldom contain links to spam pages. Some trusted pages are selected as seed set and trust scores are assigned to them, whereas the remaining pages are assigned zero trust scores. Then trust scores are propagated from seed set to all other reachable pages on the web. Hence, the pages with high trust scores are considered as good pages and those with poor trust scores are considered as spam.

3.4 Natural Language Processing Approach

This approach is based on analysis of text data of the web page. Language Analysis is performed at semantic level and syntactic level to draw various inferences. Generally, TF-IDF algorithm is used in information retrieval and text mining. TF-IDF yields a weight that measures how important a word is to a document in a corpus. The importance increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. The term frequency (TF) is simply the number of times a given term appears in a specific document. The inverse document frequency (IDF) is a measure of the general importance of the term. Roughly speaking, the IDF measures how common a term is across an entire collection of documents. TF-IDF approach has also been used to detect phishing websites [63].

In [3], authors considered a number of content based heuristics like number of words in a page, average length of words, number of words in the page title, amount of anchor text etc. to construct a decision tree classifier for spam detection. In [36], authors used divergence in language models to detect link spam in blogs on the basis that spammers create links among sites having no semantic relation.

In [53] the authors extracted features based on sentence-level topic information. They first created LDA [11] with a ham corpus and apply it to the unseen documents to infer the topic distribution of the sentences. In another work [13], authors analyzed the content of a web page and determined how closely its title is related to its body and calculate the fraction of hidden content of a web page. Westbrook and Greene [59] performed semantic analysis of textual content for spam detection. They created a page analyser by using features like sentence length, frequency of stop-word occurrence and analysis of parts of speech.

3.5 Honeypot Based Approach

A honeypot is a kind of surveillance tool that is used to monitor the activities of intruders into the system. It is a security device whose value lies in unauthorized or illicit use of that resource [52]. It works on the principle that nobody should interact with it. Thus any interaction with a honeypot signifies unauthorized access. Honeypots can be classified as physical honeypots (dedicated machine based), virtual honeypots (virtual machine based), low interaction honeypots (that work by emulating some services and operating system) and high interaction honeypots (with real operating system and applications). Honeypots have been deployed by some researchers for purpose of spam detection [37].

Anagnostakis et al. [10] designed the shadow honeypots that could identify the suspicious traffic and diverted it to a shadow version of the application. Honey Client is an active client honeypot which is used to detect browser based attacks. Moshchuk et al [38] used a virtual machine based honeypot to identify the malicious executable. They used a state- change approach that dealt with time bombs, pop-up windows, and other browser based attacks. Strider [35] is a Honey Monkey system to identify spammers by analyzing redirections on web pages. It is a high- interaction, virtual honeypot that filters false positives from the list of URLs by visiting all URLs when a browser is launched. Redirections are recorded and the most popular destination domains after redirection are marked as spam sites. Honeypot approach has also been used to design software that could automatically patch software vulnerabilities generating spam [46]. Provos et al [41] identified malicious URLs by deploying a virtual machine and taking into consideration the various factors like security, user contributed content, advertising, and 3rd-party widgets. They observed the state changes in the file system, registry, and the spawned processes in order to identify harmful URLs.

3.6 Statistical Approach

This approach explores the distribution of various properties of data sets in consideration. It assumes that outlier values detected in the distribution graphs are actually referring to spam pages. Many authors have used statistical approach to detect spam. Cafarella and Cutting [15] used statistical distribution of the words in web pages to combat the spamming techniques like adding irrelevant or repeated words

with good-looking text. Researchers found that the URLs of spam pages have exceptional number of dots, dashes, digits and length. Most spam pages that reside on the same host have a very low word count variance. Fetterley et.al [23] developed techniques for detection of phrase level stitching by performing sentence-level synthesis of web pages that consist of an unusually large number of popular phrases. They employed a technique called shingling, where they created a feature set of k-word phrases uniformly at random from each document and compared for different documents. Fetterly et al. [24] explored various features of link structure, page content, page evolution etc. and on performing statistical distribution concluded that outliers in the statistical distribution of these properties are marked as web spam.

3.7 Signature Based Approach

This approach works on the basis of known pattern of bytes that may appear in the data traffic. This system compares the incoming or outgoing data with a code considered as signature. If the match occurs, it is an indication of spam. This approach is quite simple and has been adopted by many researchers in spam detection. In [42] authors have designed a system called Spam Campaign Assassin (SCA) using signature based approach. In another work [64], hash value information (signature) was captured in the form of regular expressions based on URLs of the exploit server. The signatures so generated were distributed to browsers for detecting drive by download attacks. The work described in [28] makes use of URL properties to detect malicious pages, without requiring the crawling of pages. In the first step, each URL is analysed to find keywords delimited by common separators such as + and – and URL prefix is extracted. Then suspicious URLs are clustered to create a group of malicious links from the same campaign. Finally, for each malicious group, regular expression signatures are generated.

3.8 Fuzzy Logic Approach

This approach considers the fuzzy boundaries i.e. where a membership of a class is not concrete like true or false rather degree of truth is measured. This approach allows partial membership in a set. The degree of relevance is measured and is associated with each membership of a fuzzy set. Such systems are more suitable in situations where there is a degree of uncertainty involved. In [22], trainable fuzzy logic classifier has been used to classify e-mails into spam and ham. Their system learns various fuzzy rules at the time of training and then the inference engine classifies all the messages based on the generated rules. Fuzzy Approach has also been used in detection of phishing activities [2], phishing being one major form of spam. Authors have combined the fuzzy logic with data mining algorithms to design an anti-spam filter. Fuzzy logic has also been used in [56] for spam detection which could enhance the performance tuning too.

3.9 Biologically inspired Approach

This approach refers to artificial intelligence techniques inspired by the way in which natural systems work. It is based on the biological evolution and follows the steps of mutation, recombination, and selection. A fitness function is determined and applied on all candidate solutions and finally the optimum solution is obtained. It is one of the emerging approaches being used for spam detection. Few researchers have worked in this area.

In [34], authors have applied Artificial Neural Networks (ANN) for phishing detection which is one form of web spam. They considered 27 parameters classified them into six groups. These groups were used to train the ANN for detecting phishing websites. In [44] also authors designed Ensemble classifier taking ANN as one of the six machine learning methods used in it. The Ensemble classifier was able to enhance the accuracy in detecting spam to a great extent. Genetic Approach has also been successful in detecting spam e-mails [45] by generating spam mail prototypes. Dudley et al. [20] also used an Evolutionary Algorithm that involved a set of weights applied on Spam Assassin in order to minimize the number of false positives and false negatives. Zhang et al.[65] also proposed an evolutionary algorithm of feature selection for designing spam detection mechanism.

3.10 User Behavior Approach

Since user behavior is also good source of ranking signal [3], Liu et al. [30] proposed a number of user-behavior features for separating spam pages from normal pages. They also presented a framework that combines machine-learning techniques assisted by user behavior to detect spam pages. The spam-detection approach of Liu et al. [31] is neither content-based nor link-based, instead, it relies on user-behavior and Bayes learning. The proposed method analyses user-behavior patterns as shown in a collected Web-access log and uses three different features—search engine oriented visiting ratio, the number of clicks on hyperlinks in a Web document, and the number of sessions in a user visit.

4. DISCUSSIONS

Among the approaches discussed, machine learning approaches give results with more accuracy. However, the classifier needs to be trained with the new data when some new kind of spam is observed. While the accuracy of Artificial Neural Networks (ANN) classification is high, this approach has a tendency to require significant computing time for spam classification. However, in machine learning approaches the balance between having an up-to-date training dataset and resources available to train or retrain the artificial neural network is critical. This requires a continued effort of striking a balance between functionality and effectiveness. SVM approaches are even more computationally expensive, which to constrains their maximum potential application in online implementations.

The signature-based approaches are quite simple and able to operate online in real time. The drawback is that this approach can only detect known attacks with the predefined signatures that are produced by experts. This defence system in this approach should be able to generate the signals automatically and flexible enough to accommodate the dynamic data traffic otherwise the attackers can cause the small modifications while replicating the worms in the attack programs to circumvent the entire defence system.

Honeypot approach performs well in reducing false positives, the major benefit being the easy customization for different environments and threats. But at the same time, they involve complex deployment and maintenance. They can be very time intensive and expensive because special hardware is required for different operating systems. However honeypots are successful in capturing the traffic passing through them which can also cover new tactics or modified codes that are not used previously. In graph based solutions there are some inherent long correlations among the data objects which have great dependency on what kind of objects they are taking into consideration.

Evolutionary approach is successful in spam detection because of its versatility, robustness and adaptability. This approach is well suited in cases of uncertainty with less dependence on domain knowledge. It can easily handle large number of candidate solutions. Though this approach is capable of changing in response to environmental changes but it may involve huge computation time in encoding /decoding task. Also, genetic algorithm may fail at local optimum due to lack of hill-climbing capacity.

5. CONCLUSIONS

In this paper, we presented a comprehensive survey of various approaches used in web spam detection. We observed that there is an on-going battle between search engines and spammers. With every new design of spam detection we find spammers trying to circumvent it through its ever evolving new tactics. Due to great impact of spam on search engines and online community, web spam detection has become a key area of research in both academia and industry. In this paper, we first discussed the general phenomena of spam as challenging issue for search engines, showing the numeric estimates as provided by various security reports. We presented a brief overview of various forms of spam and discussed various approaches used for web spam detection inclusive of traditional as well as emerging approaches, presenting their underlying characteristics.

We observed that the above defined approaches have been successful in spam detection in different scenarios. There are situations where one approach outperformed the other. Each approach has its pros and cons. There is a cost attached in implementation of each approach which has to be leveraged with the success rate achieved by it in detecting spam. One has to strike a trade-off between success rate and cost. Also, the approaches could be combined to achieve excellent accuracy in spam detection.

We found that not much work has been done in spam detection using fuzzy based approach. A fuzzy approach could be used in combination with classification techniques to bring out an effective solution for spam. Evolutionary approach is still in its infancy in arena of spam detection. Genetic algorithms being an emerging area could be explored and applied into spam detection. Hence, these approaches are identified as promising directions for future research in web spam detection. Our hope is that this survey can help researchers to have insight into various techniques against spam and give them a hint for future directions in fighting against spam.

6. REFERENCES

[1] Abernethy, J., Chapelle, O., & Castillo, C. "Graph regularization methods for Web spam detection", Machine Learning, (81:2), 2010, 207-225.

[2] Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. "Intelligent phishing detection system for e-banking using fuzzy data mining", Expert systems with applications, (37:12), 2010, 7913-7921.

[3] Agichtein, E., Brill, E., & Dumais, S. "Improving web search ranking by incorporating user behaviour information", In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006 August, pp. 19-26.

[4] Akoglu, L., & Faloutsos, C. "Anomaly, event, and fraud detection in large network datasets", In Proceedings of the sixth ACM international conference on Web search and data mining , 2013, February, pp. 773-774.

[5] Almeida, Tiago A., & Akebo Yamakami. "Compression- based spam filter", Security and Communication Networks 2012.

[6] Almeida, T. A., & Yamakami, A. "Occam's razor-based spam filter", Journal of Internet Services and Applications, (3:3), 2012, pp 245-253.

[7] Almeida, T. A., & Yamakami, A. "Advances in spam filtering techniques", Computational Intelligence for Privacy and Security, Springer Berlin Heidelberg, 2012, pp. 199-214

[8] Almeida, T. A., & Yamakami, A. "Facing the spammers: A very effective approach to avoid junk e-mails", Expert Systems with Applications, (39:7), 2012, pp. 6557-6561.

[9] Amitay, E., Carmel, D., Darlow, A., Lempel, R., & Soffer, A. "The connectivity sonar: detecting site functionality by structural patterns", In Proceedings of the fourteenth ACM conference on Hypertext and hypermedia, 2003, August, pp. 38-47

[10] Anagnostakis, K. G., Sidiroglou, S., Akritidis, P., Xinidis, K., Markatos, E., & Keromytis, A. D. "Detecting targeted attacks using shadow honeypots", In Proceedings of the 14th USENIX security symposium 2005.

[11] Blei, D. M., Ng, A. Y., & Jordan, M. I. "Latent dirichlet allocation", The Journal of machine Learning research, (3:1), 2003, pp. 993-1022.

[12] Breiman, L. "Random forests", Machine learning, (45:1), 2001, pp. 5-32.

[13] Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. A. "Link-Based Characterization and Detection of Web Spam". In international workshop on adversarial information retrieval on the web , AIRWeb, 2006, August. pp. 1-8.

[14] Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. "Using rank propagation and probabilistic counting for link-based spam detection", In Proceedings of WebKDD (Vol. 6), 2006, August.

[15] Caferrella M. & Cutting, "Building Nutch: Open source search". Queue, (2: 2), 2004, pp. 54-61.

[16] Castillo, C., Donato, D., Gionis, A., Murdock, V., & Silvestri, F. "Know your neighbors: Web spam detection using the web topology", In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, July, pp. 423-430.

[17] Chang, C. C., & Lin, C. J. "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology (2:3), 2011, pp. 27-35.

[18] Cohen, W. W. & Kou, Z. "Stacked graphical learning: approximating learning in markov random fields using very short inhomogeneous markov chains" , Technical report, 2006.

[19] Dai, N., Davison, B. D., & Qi, X. "Looking into the past to better classify web spam", In Proceedings of the 5th

international workshop on adversarial information retrieval on the web, 2009, April, pp. 1-8.

[20] Dudley, J., Barone, L., & While, L. "Multi-objective spam filtering using an evolutionary algorithm". In Evolutionary Computation, IEEE World Congress on Computational Intelligence, 2008, June, pp. 123-130.

[21] Erdélyi, M., Garzó, A., & Benczúr, A. A. "Web spam classification: a few features worth more", In Proceedings of the 2011 Joint WICOW/AIRWeb ACM Workshop on Web Quality , 2011, March, pp. 27-34.

[22] Fuad, M. M., Deb, D., & Hossain, M. S. "A trainable fuzzy spam detection system", In Proc. of the 7th Int. Conf. on Computer and Information Technology, 2004, December

[23] Fetterly, D., Manasse, M., & Najork, M. "Detecting phrase-level duplication on the world wide web". In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, August, pp. 170-177.

[24] Fetterly, D., Manasse, M., & Najork, M. "Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages", In Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004 , pp. 1-6, ACM.

[25] Friedman, J., Hastie, T., & Tibshirani, R. "Additive logistic regression: A statistical view of boosting" Annals of statistics, 2000, pp. 337-374.

[26] Ghiam, Shekoofeh, and Alireza Nemaney Pour. "A Survey on Web Spam Detection Methods: Taxonomy.", arXiv preprint arXiv:1210.3131 , 2012 .

[27] Gyongyi, Z., & Garcia-Molina, H, "Web spam taxonomy", In First international workshop on adversarial information retrieval on the web AIRWeb, 2005.

[28] John, J. P., Yu, F., Xie, Y., Krishnamurthy, A., & Abadi, M. "deSEO: Combating Search-Result Poisoning" , In USENIX Security Symposium, 2011, August.

[29] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. " Design and evaluation of a real-time URL spam filtering service", In IEEE Symposium on Security and Privacy, 2011

[30] Liu, Y., Chen, F., Kong, W., Yu, H., Zhang, M., Ma, S., & Ru, L. "Identifying Web Spam with the Wisdom of the Crowds", ACM Transactions on the Web (TWEB), (6:1), 2012, pp. 2-12.

[31] Liu, Y., Zhang, M., Ma, S., & Ru, L. "User behavior oriented web spam detection", In Proceedings of the 17th international conference on World Wide Web, 2008, April, pp. 1039-1040. ACM

[32] Lu, L., Perdisci, R., & Lee, W. "SURF: detecting and measuring search poisoning", In Proceedings of the 18th ACM conference on Computer and communications security, 2011, October, pp. 467-476. ACM.

[33] Mcafe Labs Threats Report available at http://www.mcafee.com/uk/resources/reports/rp-quarterly-threat-q4-2013.pdf

[34] Martin, A., Anutthamaa, N., Sathyavathy, M., Francois, M. M. S., & Venkatesan, P. "A Framework for Predicting Phishing Websites Using Neural Networks", International Journal of Computer Science Issues, (8:2). 2011.

[35] Microsoft research strider team. Strider search defender, May 2006. http://research.microsoft.com/ SearchDefender/

[36] Mishne, G., Carmel, D., & Lempel, R. "Blocking Blog Spam with Language Model Disagreement", In In international workshop on adversarial information retrieval on the web (Vol. 5), 2005, May, pp. 1-6.

[37] Mokube, I., & Adams, M. "Honeypots: concepts, approaches, and challenges", In Proceedings of the 45th annual southeast regional conference, 2007, March, pp. 321-326, ACM.

[38] Moshchuk, A., Bragin, T., Gribble, S. D., & Levy, H. M. "A Crawler-based Study of Spyware in the Web", In NDSS, 2006, February.

[39] Najork, M. "System and method for identifying cloaked web servers", patent, 2002.

[40] Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. "Detecting spam web pages through content analysis", In Proceedings of the 15th international conference on World Wide Web, 2006, May, pp. 83-9, ACM.

[41] Provos, N., McNamee, D., Mavrommatis, P., Wang, K., & Modadugu, N. "The ghost in the browser analysis of web-based malware" , In Proceedings of the first conference on First Workshop on Hot Topics in Understanding Botnets, 2007, April, pp. 4-4.

[42] Qian, F., Pathak, A., Hu, Y. C., Mao, Z. M., & Xie, Y. "A case for unsupervised-learning-based spam filtering", ACM SIGMETRICS Performance Evaluation Review, (38:1), 2010, June, pp. 367-368).

[43] Quinlan, J. R. "C4. 5: programs for machine learning" Vol.1, Morgan kaufmann, 1993.

[44] Sanglerdsinlapachai, N., & Rungsawang, A. "Web phishing detection using classifier ensemble", In Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services, 2010, November, pp. 210-215, ACM.

[45] Sanpakdee, U., Walairacht, A., & Walairacht, S. "Adaptive spam mail filtering using genetic algorithm", Advanced Communication Technology, 2006 and ICACT 2006. The 8th International Conference (Vol. 1, pp. 441-445). IEEE.

[46] Sidiroglou, S., & Keromytis, A. D, "A network worm vaccine architecture", In Proceedings of Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises", 2003, June, pp. 220-225.

[47] Silva, R. M., Yamakami, A., & Almeida, T. A. "An analysis of machine learning methods for spam host detection", in Proceedings of 11th International Conference on Machine Learning and Applications , 2012, pp. 227-232. IEEE.

[48] Silva, R. M., Almeida, T. A., & Yamakami, A. "Artificial neural networks for content-based web spam

detection", In Proc. of the 14th International Conference on Artificial Intelligence, 2012, pp. 1-7.

[49] Silva, R. M., Almeida, T. A., & Yamakami, A. "Towards web spam filtering with neural-based approaches", In Advances in Artificial Intelligence–IBERAMIA, 2012, pp. 199-209, Springer Berlin Heidelberg.

[50] Spirin, Nikita, and Jiawei Han. "Survey on web spam detection: principles and algorithms." ACM SIGKDD Explorations Newsletter 13.2 (2012): 50-64.

[51] Sobek, M. "Pr0-google's pagerank 0 penalty. Badrank", 2002.

[52] Spitzner, L. "Honeypots: Catching the insider threat", In Proceedings of 19th Annual Conference on Computer Security Applications, 2003, December, pp. 170-179, IEEE.

[53] Suhara, Y., Toda, H., Nishioka, S., & Susaki, S. "Automatically generated spam detection based on sentence-level topic information" , InProceedings of the 22nd international conference on World Wide Web companion, 2013, May, pp. 1157-1160.

[54] Svore, K. M., Wu, Q., Burges, C. J., & Raman, A. "Improving web spam classification using rank-time features" , In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web, 2007, May, pp. 9-16. ACM.

[55] Symantec's Internet Security Threat Report http://www.symantec.com/security_response/publications/threatreport.jsp

[56] Vijayan, R., Viknesh, S. T. G. M., & Subhashini, S. "An Anti-Spam Engine using Fuzzy Logic with Enhanced Performance Tuning", International Journal of Computer Applications, (0975–8887) volume 2011.

[57] Vivekprasanth, R., and Ram Kumar P. , "Fraudulent Pages Detection Using Link Reliability And Content Based Features." In Proceedings of National Conference on Future Computing, .2012

[58] Web Sense 2013 Threat Report available at http://www.websense.com/assets/reports/websense-2013-threat-report.pdf

[59] Westbrook, A., & Greene, R. "Using semantic analysis to classify search engine spam", Class Project report at http://www. stanford. edu/class/cs276a/projects/reports.(2002-11-5).

[60] Wu, B., & Davison, B. D. "Identifying link farm spam pages". InSpecial interest tracks and posters of the 14th international conference on World Wide Web , 2005, May, pp. 820-829. ACM.

[61] Yu, H., Kaminsky, M., Gibbons, P. B., & Flaxman, A. "Sybilguard: defending against sybil attacks via social networks", ACM SIGCOMM Computer Communication Review, (36:4), 2006, pp. 267-278.

[62] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, " Combating web spam with TrustRank", Proc. of the 30th International Conference on Very Large Data Bases (VLDB), Toronto, Canada, 2004.

[63] Zhang, Y., Hong, J. I., & Cranor, L. F. "Cantina: a content-based approach to detecting phishing web sites", In Proceedings of the 16th international conference on World Wide Web, 2007, May, pp. 639-648. ACM.

[64] Zhang, J., Seifert, C., Stokes, J. W., & Lee, W. "Arrow: Generating signatures to detect drive-by downloads", In Proceedings of the 20th international conference on World wide web, 2011, March, pp. 187-196, ACM.

[65] Zhang, Y., Li, H., Niranjan, M., & Rockett, P. "Applying cost-sensitive multiobjective genetic programming to feature extraction for spam e-mail filtering", Genetic Pro gramming, Springer Berlin