2016 IEEE 28th International Conference on Tools with Artificial Intelligence

Novel Features for Web Spam Detection

Santosh Kumar
School of Electrical and Computer Science
Victoria University of Wellington
Wellington, New Zealand
Email: santosh.kumar@ecs.vuw.ac.nz

Xiaoying Gao
School of Electrical and Computer Science
Victoria University of Wellington
Wellington, New Zealand
Email: xiaoying.gao@ecs.vuw.ac.nz

Ian Welch
School of Electrical and Computer Science
Victoria University of Wellington
Wellington, New Zealand
Email: ian.welch@ecs.vuw.ac.nz

***Abstract*—Recent research on web spam detection has shown promising results, and many new and efficient detection algorithms have been developed. While most research focuses on developing algorithms, our investigation shows that the features used in the algorithms are in fact very important, and different features can lead to very different results. This paper investigates three types of web spam, content-based, link-based and cloaking, and introduces new features for identifying the three types of spam. Our experimental results show that the introduction of new features significantly improves the detection performance.**

I. Introduction

Web spam is one of the main current problems of search engines because it actively degrades the quality of the search results. Many people encounter web spam by continuously finding spam content when they look for legitimate content using search engines. Over the past few years, there has been various research published in the detection of these pseudo web pages, although in response, new spam techniques have developed by spammers.

In research [1], authors explore the trade-off between feature selection and web spam classification performance. The authors conclude that larger number of features helps classifier model to achieve better accuracy; however, appropriate choice of the classifier model is more likely to be more important than constructing new sophisticated features. Feature selection plays a vital role in improving the effectiveness of web spam detection problem. Accomplishing a reduction of the number of relevant web spam features without an adverse effect on classifier accuracy increase the available processing time of web spam detector and would reduce the required system resources. An additional goal is to be able to detect all three types of web spam features: content spam, link spam, and cloaking spam features [2]. In most of the contemporary research [2], [3], [4], feature selection depends upon expect knowledge but also can only identify one or two types of web spam.

New, highly discriminative features are required to detect all three categories (content spam, link spam, and cloaking spam) of web spam. The major problem to identify all types of web spam is to choose significant measures that can promptly discover the significance and the relationship between features of the particular dataset. Since the significance and the relationship are usually described concerning correlation or mutual information.

A. Contribution

In this paper, we propose new spam features to classify spam web pages into three different categories (content, link, and cloaking). While most previous research [2], [3], [4] using content and link-based features to detect the individual type of web spam concentrated on quantifiable features, we propose novel qualitative spam features to improve web spam detection accuracy grouped our proposed features are into three sets: 1) a group of link-based features which check the reliability of links, 2) a group of content-based features extracted from various contents of a web page, and 3) a group of cloaking-based features extracted from different web page scripts and IP address of network. Finally, we adopt a learning classifier that combines all three types of the features, achieving a precision that advances the results of each type spam separately, and those accomplished by other research. This paper has the following major contributions:

1) We propose three groups of novel, highly discriminative features that enable learning classifier to deliver a superior performance and inclination on both detection and type of spam identification of web spam; types of spam include content, link, and cloaking. Our novel features provide a much larger coverage than existing methods while maintaining a high accuracy.
2) Preliminary experiments on standard WEBSPAM-UK2007 [5], ClueWeb-2009 [6], and ECML-PKDD-2011 [7] benchmark datasets demonstrate the effectiveness of the novel features on learning the classifier for detecting web spam.

The rest of the paper is formed as follows: We review the previous research work in Section 2. In section 3, we describe the proposed groups of novel web spam features. Experimentally evaluation of the proposed novel web spam features on the standard collection is presented in section 4. At last, we give the concluding remarks and future research directions.

II. Related Work

Recent works [3], [8] on web spam detection mainly focused on three types of spam: link spam, content spam, and cloaking. In this section we review recent work on spam detection that have used the WEBSPAM-UK2006, WEBSPAM-UK2007,

2375-0197/16 $31.00 © 2016 IEEE
DOI 10.1109/ICTAI.2016.93

ClueWeb 2009, and ECML-PKDD 2011 data sets. We have focused on this work because the use of these publicly available datasets have allowed us to compare our approach with other techniques and because they represent the current state-of-the-art research. In research [9], the authors proposed ten novel spam features generated by genetic programming to improve the link-spam classification using WEBSPAM-UK2006 dataset.

In research [10], authors adopted an information gain based model and proposed ten novel content-based features selected by various feature selection methods using ClueWeb 2009 dataset. Later, Algur and Pendari [11] used the original hybrid spamicity approach to propose various unique topical diversity measures for the content spam detection. Erdelyi. M [12] investigated how powerful features increase spam classification accuracy. For content-based features the author [13] also used the Latent Dirichlet allocation language model and proposed various sets of content based features using WEBSPAM-UK2007 and ClueWeb 2009 data sets.

From our point of view, the most interesting work presented results for both content and link-based spam detection using both WEBSPAM-UK2006 and WEBSPAM-UK2007 data sets. The authors [14] introduced new features based on qualified link analysis and language models. The significant aspect of this research was the significantly better than the previous results for both types of features on both the data sets.

Based upon a review of related work, we decided to concentrate on determining how the new features might improve classifier performance before investigating new classification models. Our aim is to create such group of features that allows a classifier model to reach stable results for all benchmark data sets (WEBSPAM-UK2007, ClueWeb 2009, and ECML-PKDD 2011).

III. Novel Web Spam Features

In this section, we describe the groups of common novel features used in web spam detection for all three (WEBSPAM-UK2007, ClueWeb 2009, and ECML-PKDD 2011) benchmark data sets.

A. Novel Content-based Spam Features

We have re-analysed web page content and its properties from the literate and propose the following novel heuristic features:

1) Advanced word length: We propose an advanced word length feature, which is the average length the page content without taking into account the HTML tags or the stop words. Our reason is that HTML tags are not content that is going to be presented to the end user and would include noise in the results. Also, we have observed that non-spam pages often use a significant number of stop words. This happens since a legitimate web page contains a reasonable amount of prepositions, articles or, conjunctions. However, spam pages largely concentrate on the injection of keywords that enhance their position in the search engine rankings.
2) Particular phrases: It is common that content spam pages contain basic terms, phrases or queries from the search engines. Also, it has been observed that in other domains like email spam where emails frequently include general sentences or particular spam words. After analysing all three benchmark data sets (WEBSPAM-UK2007, ClueWeb 2009, and ECML-PKDD 2011), we extracted the specific phrases and created a list of the common phrases like ”free”, ”urgent”, ”click here link”, ”Free membership” and ”buy”.
3) Number of keywords or description terms: We have done an analysis on the number of terms or keywords utilized in the description and keywords properties of the META tag, in both spam and non-spam pages. We found in our analysis that pages with less than 120 keywords or description terms have less of 18% of probability of being spam. Moreover, if the number rises then the possibility of a web page being spam increases as well. This is the indication but not sufficient by itself. Therefore, we examined not only the description and keywords properties of the META tag but also the number of appearances of these keywords or terms in the properties and within the page content.
4) Ratio of bytes and total bytes of HTML code: Through the analysis of the web pages, we found that the ratio between the size in bytes of the HTML scripting code of a page and the total number of bytes increases according to the likelihood of spam. The longer the code in the HTML, the higher the probability of content spam.
5) Number of stop keywords: In search engine indexing process, the search engine extracts the all possible stop keywords (able, about, begin, both, can, cause, how, however, many, and may) from the web page content and then indexes it. Therefore, the spammers frequently choose to insert garbage content, without prepositions, articles, conjunctions, or frequent extra keywords without using stop keywords. We propose to analyze the number of stop keywords in the page content.

B. Novel Link-based Spam Features

1) Number of broken links: Broken links are a well-known problem for both spam and non-spam web pages. Besides, these broken links have an adverse impact on PageRank. In our analysis, a web page contains the higher of broken links has more probability to be a spam web page.
2) Anchor text typology: Typically spam web pages contain automatically generated text and links. Furthermore, the anchor text of many links has commonly created the thought in the context of the search engines rather of its users. Therefore, we have chosen four features in order to measure the number of links that are created only by 1) punctuation lines, 2) numbers, 3) a URL, and 4) an uninhabited string.
3) Unusual combinations in the link: In our analysis we identify unusual combination in the link. The unusual combination can be defined as a string of lower-case

letters with any upper-case letters or digits between segments of the string, i.e., Goog1e, credit4U, Free4u, Yah0o. These types of combination we extracted with the help of regular expression. In particular, we calculated the ratio of total appearance of unusual combinations within the different domains and the degree of unusual combinations appearance in the visible text of a link or page.

4) Parked domains: The web host relates to a domain that is only a parking website with no real content but many links. There is usually a plan for bidding for the domain name and many links to other domains held by the same operator.
5) Mutual link counts in web graph: Additionally to the user-URL bipartite graph, we recommend examining the link structure between the in-links and out-links to distinguish either link exchanges have indeed come into presence. For this purpose, we use of a web graph obtained from a commercial search engine.
6) Combination of link attributes: In our analysis we combine and calculated common score for various attributes of a link: link creator, page type, link context, link placement, link color and link relationship.

C. Novel Cloaking Spam Features

1) Request URL: We use request URL to indicate the external objects (i.e; images, external scripts, CSS) that are loaded from some other URLs or resources.
2) Suspicious URL: The URL of phishing website may accommodate the legitimate web-site's URL as a substring $(http://www.yahootag.com)$, or may be similar to the legitimate URL $(http://www.yaho0.com)$ in which the letter o in yahoo is substituted with number 0. Sometime the IP addresses are used to mask the host name $(http://255.255.255.255/google.html)$
3) Suspicious JavaScript: Malformed Javascript tags may indicate that the page includes scripts that are intended to send personal information or device information to phishers.
4) Link Redirection: In this feature, we record the number of times that a particular URL redirected to a different URL, for example, by response with HTTP status 302, or by the setting of specific JavaScript properties. The number of redirection represents the number of domain linked to a particular URL.
5) JavaScript Injection: In some cases, malware authors insert malicious JavaScript code into existing benign scripts on a compromised host. In particular, it is common for malware authors to add their malicious scripts to the code of popular JavaScript libraries hosted on a compromised site. We calculate the number of jQuery and SWFObject functions used in particular web-page at runtime.
6) Sequences of Method Calls: In this feature we monitor the sequence of method invocations on instantiated and ActiveX controls.
7) Repeated Pattern: We also observed that, in certain cases, an Abstract Syntax Tree (AST) may contain a set of nodes repeated a large number of times. This commonly occurs when the script uses some JavaScript data structure that yields many repeated AST nodes. For example, malicious scripts that unpack or de-obfuscate their exploit payload frequently utilize a JavaScriptArray of elements to store the payload.
8) Similar Code: We identify blocks of code that are shared across two scripts, and it can be the case that these blocks are not continuous. One script can be broken down into small fragments that are matched to the other script in different positions. This is why we take into account the fragmentation of the matching blocks.
9) Suspicious Tokens: The suspicious token model determines if the values of a certain feature are elements of an enumeration, i.e., token are drawn from a limited set of alternatives. In legitimate scripts, certain features can often have a few possible values. For example, in an ActiveX method that expects a Boolean argument, the argument values should always be 0 or 1. If a script invokes that method with a value of $0x0c0c0c0c$, the call should be tagged as anomalous. We apply this model to each method parameter and property.

IV. Experiments and Results

A. Significance of proposed features

We use a classification tree method to evaluate the significance of proposed novel web spam features (content, link, and cloaking). This classification tree method chooses the features based on the Gini coefficient, which is a measure of the statistical distribution among the features [15]. The significance of particular feature is calculated by summing over all nodes in the tree, the drop in node impurity. We estimated the significance of novel features for each data set separately. In Table 1, we report the average significance (rank and values) of the novel features (other features are not shown).

B. Data sets & Methodology

We used three benchmark data sets (WEBSPAM-UK2007 [5], ClueWeb-2009 [6], and ECML-PKDD-2011 [7]) in our experiments. In this section, we present the learning classifier model used as well as the test method classifier performance and the validation of results evaluation for web spam detection. In our experiments, we use Dual-Margin based Multi-Class Hypersphere Support Vector Machine (DMMH-SVM) as the learning classifier model and a basic Multi-Class Support Vector machine [16]. The focus of this paper is the impact of choice of features on results so we refer you Santosh [17] for technical specifications.

C. Methods of tests and evaluations & Influence of novel features

Six times 5-fold cross-validation [18] is run on the data sets to estimate performance. The precision, recall, true positive

Rank	Feature	Average Value	Significance	Rank	Feature	Average Value	Significance
1	Advance word length	3.2	3.08E-07	11	Similar code	3.2	7.88E-09
2	Particular phrases	3.2	2.40E-07	12	Suspicious tokens	3.3	5.27E-11
3	Request URL	3.3	2.22E-07	13	Unusual combinations in the link	3.4	5.31E-11
4	Suspicious URL	3.5	2.06E-07	14	Parked domains	3.6	5.25E-11
5	Broken link	3.4	1.42E-07	15	Number of stop keywords	3.5	5.74E-11
6	Anchor text typology	3.3	1.30E-07	16	Suspicious JavaScript	3.2	5.82E-11
7	Number of keywords or description terms	3.1	9.90E-09	17	Link redirection	3.3	5.62E-11
8	Ratio of bytes and total bytes of HTML code	3.6	9.32E-09	18	JavaScript injection	3.2	5.29E-11
9	Sequences of method calls	3.4	8.68E-09	19	Mutual link counts in web graph	3.4	5.94E-11
10	Repeated pattern	3.3	8.48E-09	20	Combination of link attributes	3.1	5.86E-11

TABLE I
THE AVERAGE SIGNIFICANCE OF THE NOVEL FEATURES

rate(TP), false positive rate(FP), accuracy, and F-measure are used to measure the performance. To demonstrate the influence of our novel features, we compare our results with existing features mentioned in [3], [19], [20]. To prove that difference in results are significant, we performance the test on two difference multi-class support vector machines: MSVM [16] and DMMH-SVM [17]. Table 2 presents the results for MSVM learning classifier with existing and novel features on three benchmark data sets (WEBSPAM-UK2007:**A** [5], ClueWeb-2009:**B** [6], and ECML-PKDD-2011:**C** [7]). Where Table 3 presents the results for DMMH-SVM learning classifier model with existing and novel features on the three data sets.

D. Results discussion

The contemporary research [2], which are on WEBSPAM-UK2007:**A** , ClueWeb-2009:**B** , and ECML-PKDD-2011:**C**, are the best reference to compare with our results. The average accuracy (among all spam classes) for M-SVM classifier model with precomputed features obtained in our paper is 80%, 82%, and 80% for data sets WEBSPAM-UK2007, ClueWeb-2009, and ECML-PKDD-2011 respectively. Table 2 demonstrate that with same classifier model our novel features achieved significant results and the average accuracy increases to 87%, 88%, and 86% for data sets WEBSPAM-UK2007, ClueWeb-2009, and ECML-PKDD-2011 respectively.

To demonstrate that the results achieved by our novel features are statistically significant, we choose a different learning classifier model (DMMH-SVM) to test the performance of our features for all three mentioned data sets. Table 3 demonstrates that our novel features achieved even better accuracy and F-measure score in compare with the previous classifier model. However, the ratio or the percentage between results for pre-calculated and novel features are stable for sensitivity, specificity, accuracy and F-measure matrices.

The F-measure obtained in Ntoulas [2] for the WEBSPAM-UK2007 (content spam), data set was 82 % and the result for Link-spam was 80 %. With the help of novel content and link-based features, we achieve better 97% and 98% results for content and link-spam detection respectively. The specificity and the sensitivity are not stable in research [21] , which the adopted set of parameters for their results evaluation. Table 1 demonstrate that due to our novel features, we manage to achieve a balance threshold (between sensitivity and the specificity) to detect the all three spam classes.

V. CONCLUSION

This paper details the novel features we used for web spam detection. Our experimental results show that the detection performance improves significantly with our novel features. A total of 20 new features are introduced and they improved the average detection rate by 97%. In the future work, we will investigate other type of spam, especially new spam to further expand the features. We are also developing new classification algorithms to better utilize these features for further improving the performance. Our recent research on applying transfer learning methods on Web spam detection has shown promising results. We will further study the feature distributions and common feature characteristics, and to develop a system that learns the common knowledge for transfer learning.

REFERENCES

[1] M. Erdélyi, A. Garzó, and A. A. Benczúr, "Web spam classification: a few features worth more," in *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality*. ACM, 2011, pp. 27–34.

[2] N. Spirin and J. Han, "Survey on web spam detection: principles and algorithms," *ACM SIGKDD Explorations Newsletter*, vol. 13, no. 2, pp. 50–64, 2012.

[3] A. Taweesiriwate, B. Manaskasemsak, and A. Rungsawang, "Web spam detection using link-based ant colony optimization," in *Advanced Information Networking and Applications (AINA), 2012 IEEE 26th International Conference on*. IEEE, 2012, pp. 868–873.

[4] R. M. Silva, T. A. Almeida, and A. Yamakami, "Artificial neural networks for content-based web spam detection," in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012, p. 1.

[5] D. D. Carlos Castillo. (2008) Web spam challenge 2007/2008 @ONLINE. [Online]. Available: http://chato.cl/webspam/

[6] G. V. Cormack. (2009) Waterloo spam rankings for the clueweb09 dataset @ONLINE. [Online]. Available: https://plg.uwaterloo.ca/ gvcormac/clueweb09spam/

[7] N. Antulov-Fantulin, M. Bošnjak, M. Znidaršic, M. Grcar, M. Morzy, and T. Šmuc, "Ecml-pkdd 2011 discovery challenge overview," *Discovery Challenge*, pp. 7–20, 2011.

Results with existing features					Results with novel features				
Spam-class	Metric	A	B	C	Spam-class	Metric	A	B	C
Content	Sensitivity	77	80	80	**Content**	Sensitivity	86	87	86
	Specificity	84	82	78		Specificity	92	90	89
	Accuracy	83	85	84		Accuracy	88	90	88
	F-measure	82	84	83		F-measure	86	89	88
Link	Sensitivity	75	82	80	**Link**	Sensitivity	87	88	87
	Specificity	82	83	85		Specificity	90	90	88
	Accuracy	82	84	84		Accuracy	88	90	86
	F-measure	80	85	84		F-measure	87	89	88
Cloaking	Sensitivity	78	83	82	**Cloaking**	Sensitivity	89	86	90
	Specificity	84	83	84		Specificity	88	89	88
	Accuracy	84	83	82		Accuracy	89	88	87
	F-measure	82	85	83		F-measure	88	88	88

TABLE II
MULTI-CLASS SPAM CLASSIFICATION (M-SVM) RESULTS WITH EXISTING AND NOVEL SPAM FEATURES (%)

Results with existing features					Results with novel features				
Spam-class	Metric	A	B	C	Spam-class	Metric	A	B	C
Content	Sensitivity	89	82	82	**Content**	Sensitivity	**96**	**97**	**97**
	Specificity	89	87	88		Specificity	**97**	**96**	**97**
	Accuracy	86	87	87		Accuracy	**97**	**96**	**96**
	F-measure	84	86	85		F-measure	**97**	**97**	**96**
Link	Sensitivity	85	82	80	**Link**	Sensitivity	**97**	**96**	**97**
	Specificity	88	92	89		Specificity	**97**	**96**	**97**
	Accuracy	88	86	87		Accuracy	**97**	**96**	**97**
	F-measure	87	89	88		F-measure	**98**	**96**	**97**
Cloaking	Sensitivity	83	86	85	**Cloaking**	Sensitivity	**97**	**96**	**97**
	Specificity	88	89	88		Specificity	**95**	**96**	**96**
	Accuracy	88	86	87		Accuracy	**97**	**96**	**96**
	F-measure	86	87	88		F-measure	**96**	**98**	**97**

TABLE III
MULTI-CLASS SPAM CLASSIFICATION (DMMH-SVM) RESULTS WITH EXISTING AND NOVEL SPAM FEATURES (%)

[8] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[9] L. Shengen, N. Xiaofei, L. Peiqi, and W. Lin, "Generating new features using genetic programming to detect link spam," in *Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference on*, vol. 1. IEEE, 2011, pp. 135–138.

[10] M. Mahmoudi, A. Yari, and S. Khadivi, "Web spam detection based on discriminative content and link features," in *Telecommunications (IST), 2010 5th International Symposium on*. IEEE, 2010, pp. 542–546.

[11] S. P. Algur and N. T. Pendari, "Hybrid spamicity score approach to web spam detection," in *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on*. IEEE, 2012, pp. 36–40.

[12] M. Erdélyi, A. A. Benczúr, J. Masanés, and D. Siklósi, "Web spam filtering in internet archives," in *Proceedings of the 5th international workshop on Adversarial information retrieval on the web*. ACM, 2009, pp. 17–20.

[13] I. Bíró, D. Siklósi, J. Szabó, and A. A. Benczúr, "Linked latent dirichlet allocation in web spam filtering," in *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*. ACM, 2009, pp. 37–40.

[14] L. Araujo and J. Martinez-Romo, "Web spam detection: new classification features based on qualified link analysis and language models," *Information Forensics and Security, IEEE Transactions on*, vol. 5, no. 3, pp. 581–590, 2010.

[15] G. Behera, "Privacy preserving c4. 5 using gini index," in *Emerging Trends and Applications in Computer Science (NCETACS), 2011 2nd National Conference on*. IEEE, 2011, pp. 1–4.

[16] A. Mathur and G. Foody, "Multiclass and binary svm classification: Implications for training and classification users," *Geoscience and Remote Sensing Letters, IEEE*, vol. 5, no. 2, pp. 241–245, 2008.

[17] I. W. M. M. Santosh Kumar, Xiaoying Gao, "A machine learning based web spam filtering approach," in *Advanced Information Networking and Applications (AINA), 2016 March The 30th IEEE International Conference on*. IEEE, 2016.

[18] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

[19] B. Wu and B. D. Davison, "Identifying link farm spam pages," in *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, ser. WWW '05. New York, NY, USA: ACM, 2005, pp. 820–829.

[20] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006, pp. 83–92.

[21] K. L. Goh, A. K. Singh, and K. H. Lim, "Multilayer perceptrons neural network based web spam detection application," in *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on*. IEEE, 2013, pp. 636–640.