OLAWSDS: An Online Arabic Web Spam Detection System

Mohammed N. Al-Kabi
Faculty of Sciences & IT
Zarqa University
Zarqa, Jordan

Heider A. Wahsheh
Computer Science Department
College of Computer Science
King Khalid University
Abha, Saudi Arabia

Izzat M. Alsmadi
Information Systems Department
College of Computer & Information
Sciences Prince Sultan University
Riyadh 11586, P. O. Box 66833,
Saudi Arabia

***Abstract*—For marketing purposes, Some Websites designers and administrators use illegal Search Engine Optimization (SEO) techniques to optimize the ranking of their Web pages and mislead the search engines. Some Arabic Web pages use both content and link features, to increase artificially the rank of their Web pages in the Search Engine Results Pages (SERPs).**

This study represents an enhancement to previous work in this field. It includes the design and implementation of an online Arabic Web spam detection system, based on algorithms and mathematical foundations, which can detect the Arabic content and link web spam depending on the tree of the spam detection conditions, beside depending on the user's feedback through a custom Web browser. The users can participate in making the decision about any Web page, through their feedbacks, so they judge if the Arabic Web pages in the browser are relevant for their particular queries or not. The proposed system uses the extracted content and link features from Arabic Web pages to determine whether to label each Web page as a spam or as a non-spam. This system also attempts to learn from the user's feedback to enhance automatically its performance.

Statistical analysis is adopted in this study to evaluate the proposed system. Statistical Package for the Social Sciences (SPSS) software is used to evaluate this new system which considers the users feedbacks as dependent variables, while Arabic content and links features on the other hand are considered independent variables. The statistical analysis with the SPSS is used to apply a variety of tests, such as the test of the analysis of variance (*ANOVA*). *ANOVA* is used to show the relationships between the dependent and independent variables in the dataset, which leads to solving problems and building intelligent decisions and results.

Keywords—Arabic Web spam; content-based; link-based; Information Retrieval

I. Introduction

Arab Internet users suffer from two problems, the first problem is the low percentage of the Internet Arabic content, and the second problem is Arabic Web spam which leads Web search engines to refer to irrelevant Web pages. The success of spamming techniques to deceive a search engine leads the Internet users to lose credibility in the search engine they used, in addition to some other negative aspects of spamming such as wasting the time and efforts of the search engine users.

This study proposes an integrated system to reduce the Arabic content and link Web spam, and filter the search engines from these malicious Arabic web pages. Although this study relies on a set of content and link Arabic Web spam conditions that have been used before, however this study differs from its predecessors by involving the Web search engine users to assess the relevancy of Arabic Web pages rendered by Search Engine Results Pages (SERPs).

The proposed system allows users to use a synchronization technique, in which the users can browse the Arabic Web pages, and give their feedbacks assessment for each visited Web page under some security considerations and confidentiality. The use of a synchronization technique helps the proposed system to ensure that the submitted assessment is conducted by users not agents and robots.

The evaluation of the results of the proposed system is based on the use of Statistical Package for the Social Sciences (SPSS) software, which enables us to conduct a statistical analysis, and confidence predictive method. SPSS software considers Arabic Web spam features as independent variables, while it considers the Search Engine Ranking (SER), TrustRank, and link popularity scores as dependent variables. The statistical analysis in SPSS applies a variety of tests, such as the test of the analysis of variance (*ANOVA*). *ANOVA* has two types (one-way and two-way analysis of variance). In this study we used two-way analysis of variance to show the relationships between the dependent and independent variables in the dataset.

The main aim of this research is the development of a system which can filter the search engines from unwanted and spam Web pages based on the Web pages' features and the users which have a main role in determining the relevancy of SERPs with their different queries.

The rest of the paper is divided as follows: Section two presents selected related work of Web spam studies. Section three presents developed system overview. Section four elaborates experiments and results. Section five summaries the paper and its contribution.

II. Related Works

The literature is rich with many studies related to Web spam, where this topic is studied from different perspectives. This section presents few of these studies which are closely

This research is funded by the Deanship of Research and Graduate Studies in Zarqa Private University / Jordan.

related to the subject of this paper: Detection of Web spam, and those studies dedicated to the evaluation of the correlation between spam and the trust.

The authors of this study enhanced the previous study of [1], which built Arabic content/link Web spam detection system. The study of [1], collected a large data set of Arabic Web pages (spam and non-spam), where various number of content and link based features extracted. Arabic content/link Web spam detection system based on the tree of the decision tree machine learning algorithm to build the rules of the proposed system, which yields the accuracy of 90.10% for Arabic content based, 93.10% for link based, and 89.01% in detecting both Arabic content and link Web spam detection system.

Content trust is essential to determine the quality of Web content, and a hot topic of research. The task of determining trustworthy information from inaccurate or untrustworthy information is becoming a hard task. The study of [2] shows how to adopt content trust to detect Web spam and rank each Web page accordingly. Text feature attributes, and information quality are used their novel content trust learning algorithm. They also developed a system to detect Web spam which shows its effectiveness relative to other alternative ways.

The study of [3] presents a new ranking algorithm for Web search engines which capable to eliminate spam Web pages from their results. A small blacklist of classified spam web pages is used by their algorithm. This ranking algorithm is based in its identification of spam on two aspects tendency and authority. Therefore a high quality Web pages with low or no spam tendency will get high ranks by their ranking algorithm. They conducted tests which show the effectiveness of this ranking algorithm relative to PageRank algorithm.

The paper of [4] studied the feedback of the users and converted it to the query log. For each user, a query log file was assigned. This log file contains: query words, document returned to the search engine, Web documents that users triggered within clicked date and time, and the rank of retrieved documents. The researchers applied two approaches: Web spam detection, and query spam detection. Web spam detection removes spam link and content features from the query log graphs, while query spam detection eliminates all queries that gain a high number of spam Web pages.

In [5] a language model approach was proposed, which extracted a combination of content-based and link-based features from two popular spam datasets (Webspam-UK2006 and Webspam-UK2007). Kullback-Leibler (KL) divergence was applied on the spam Web pages to characterize the relation between the two linked Web pages. The proposed model has improved the F-measure of Webspam-UK2006, and Webspam-UK2007 to about 6% and 2% respectively.

The study of [6] presented the influence of cloaking techniques to increase the rank of Web pages. Lin study proposed three techniques: TagDiff2, TagDiff3, and TagDiff4 to determine if the URLs are cloaked [6]. The proposed techniques are based on discovering differences in the copies (HTML tags) of a specified Web page when it is sent to the Web crawler and to the Web browser. The conducted tests showed that tag-based methods exceed the link-based and content-based results in precision and recall. The Decision tree J48 uses the integration of content-based and tag features to yield an accuracy of 90.48%.

The study of [7] proposed a new methodology to detect spam Web pages based on the Qualified-Link (QL) analysis, and content-based features with the language-model (LM). Kullback-Leibler (KL) divergence was applied on the spam Web pages to find the relation between two linked Web pages based on both the content-based and link-based features. An automatic classifier was built to combine QL and LM features. The conducted results were applied on WEBSPAM-UK2006, and WEBSPAM-UK2007 datasets and showed an accuracy of 89.4% and 54.2% respectively.

Cloaking is a known Web spam technique which is used to deceive Web search engines, where the content of a Web page presented to Web search engine crawlers is different from the content of a web page provide to a Web browser. Therefore the study of [8] presents three tag-based methods to identify cloaked URLs. The effectiveness of their methods is compared against the effectiveness of term- and link-based methods, and the results prove that these three tag-based methods are more effective than term- and link-based methods. Also he presents in his paper a taxonomy to classify various cloaking detection methods. Lin study described and discussed dynamic cloaking.

The combined usage of trust and distrust propagations by semi-automatic anti-spam algorithms proved its effectiveness, but little work is done in this field. Therefore the study of [9] presents a framework to assign for each Web page a GoodRankscore (trustworthy score) and BadRank score (untrustworthy score). Afterward they propose a novel Good-Bad Rank (GBR) algorithm, where the propagation of a page's trust/distrust is based on probability of the Web page being trusted/distrusted. Tests conducted by those researchers show the effectiveness relative to link-based anti-spam algorithms that propagates only trust or distrust.

III. System Overview

This study aims to develop and improve the techniques used in [1], and proposed new system called Online Arabic Web Spam Detection System (*OLAWSDS*).

The authors of [1] built an Arabic content/link Web Spam Detection System, which mainly consists of the following main parts:

1) Built in Web crawler: The role of the crawler to automate fetching the content of Arabic Web pages.

2) Arabic Web Spam data collections: It contains 23,000 labeled Arabic spam and non-spam Web pages.

3) Arabic content/link Web pages Analyzer: This is a customized tool that analyzes and measures content and links of Arabic Web pages features, in order to evaluate their optimized features. Table 1 summarizes the main extracted content/link features.

TABLE I. ARABIC CONTENT/LINK WEB SPAM FEATURES [1].

Arabic Content Web spam features	Arabic Link Web spam features
1. Meaningless keyword stuffing (Arabic/English/Symbol) (in Web pages, Meta tags).	2. Number of image links.
3. Compression ratio for Web pages.	4. Number of internal links.
5. Number of images.	6. Number of external links.
7. Average length of Arabic/English words inside the Web pages.	8. Number of redirected links.
9. URL lengths.	10. Number of empty link text.
11. Size of compression ratio (in Kilobytes).	12. Number of empty links.
13. Web page size (in Kilobytes).	14. Number of broken links (which refers to null destinations).
15. The maximum Arabic/English word length.	16. The total number of links (the internal and external).
17. Size of hidden text (in Kilobytes).	
18. Number of Arabic/English words inside <Title tag>.	

The authors of [1] labeled the Web pages as either spam or non-spam pages in the data collections depending on their judgments and previous Arabic/ non-Arabic Web spam studies.

In this study we improved the Arabic Web Spam Detection System, by computing more derived features and benefits of the client/server model.

OLAWSDS computes the following derived features:

- Search Engine Ranking score.
- TrustRank score.
- Link popularity score.

Client/server model is a distributed system architecture, which divides the work between the server that provides the hosting of the services, and the clients which request the services. Our proposed *OLAWSDS* used the client/server model, by considering the improved Arabic Web Spam Detection System as a server. *OLAWSDS* built a custom Web browser used to explore the Web pages through the clients' computers, and the clients used it as judgments area by including their decisions, which send to the server. Figure 1 presents the main parts of *OLAWSDS* and the flow of work.

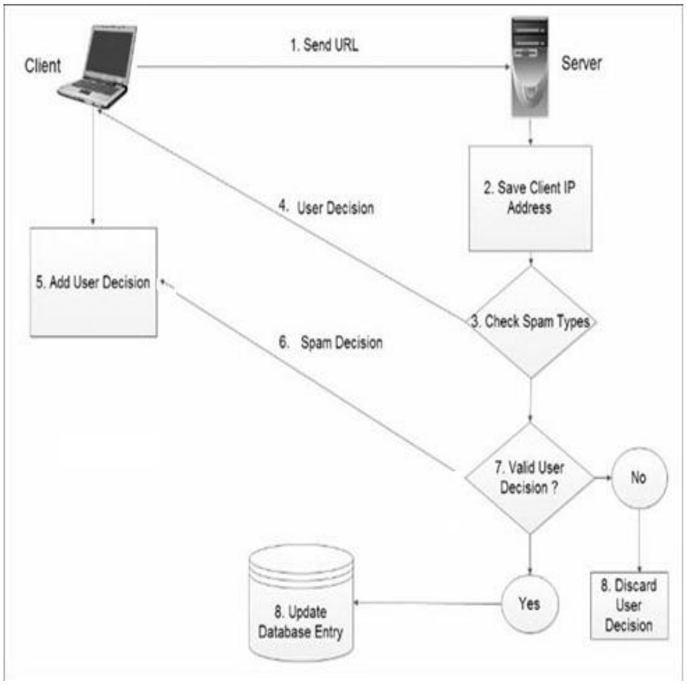

Fig. 1. OLAWSDS Architecture

In the proposed *OLAWSDS* system, we used the Arabic Web spam dataset of [1] as a black list dataset. Every user in client using *OLAWSDS* can browse the Web pages, and check and identify any spammed Web page(s).

The clients send their feedbacks to *OLAWSDS*, which reside in the server. *OLAWSDS* considered the valid users decision when their metrics exceed the thresholds. The thresholds depend on how many users send their decisions for particular Web page(s). *OLAWSDS* system saves the client server IP address and computes the Web spam features including the user's decision features. *OLAWSDS* sends the user final decision about a particular Web page either as a spam or non-spam, then update the Arabic Web spam black list dataset.

Visual and audio code is used by *OLAWSDS* to avoid spammers and robots from getting into this system, and make a fake decision of the type of Web pages. To avoid the participant of the spammers or the robots as user's clients, *OLAWSDS* requests from any client before sending the decision to fill the visual or audio code. Figure 2 presents an example of used visual and audio code.

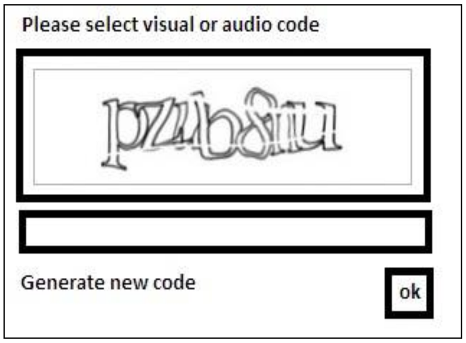

Fig. 2. Example of used Visual and audio code

OLAWSDS also saves the client IP address and prevent any decision from the same user for the same Web pages for one day.

IV. Experiment And Results

In this study we used the SPSS software to evaluate our proposed *OLAWSDS*. Analysis of Variance (*ANOVA*) was computed for 10,000 Arabic Web pages.

Statistical Package for the Social Sciences (SPSS) is a Statistical Software Package acquired by IBM in 2009, used for data mining, text analytics, and statistical analysis. SPSS divides the variable in the dataset to the main two types; dependent and independent variables, then applies the regression analysis [10].

- Dependent variables: It is defines as the results, operational outputs of the independent variables [11].

In this study, our proposed *OLAWSDS* computes the search engine ranking score. Which considered as dependent variables, where the computed score depends on the content and link features of the collected Web pages.

OLAWSDS also computes the TrustRank score, which is a link analysis technique which identify useful Web pages which are linked in most cases to other good Web pages, while spam Web pages point to the spam Web pages [12]. As search engine ranking score, the TrustRank depends on the many features of content and links features it is considered as a dependent variable.

OLAWSDS computes the link popularity score based on both external and internal links, so link popularity score considered as dependent variable.

- Independent variables: It is defined as the inputs of the independent variables (the values that determine the values of other variables) [11]. All Arabic web spam features are considered as independent variables.

ANOVA has two main types (one-way and two-way analysis of variance). It is used to show the relationships between the dependent and independent variables in the dataset. In this study we used *ANOVA* with two-way analysis of variance to determine the effect of independent variables on two or more continuous dependent variables [11].

The null hypothesis is that there is no relationship between two measured variables, or that the independent variable has no effect on the dependent variables (i.e. means are same). The alternative hypothesis states that there is relationship and effect between two measured variables (i.e. means are different). The goal of *ANOVA* is to accept or reject the null hypothesis [11].

We applied two-way analysis of variance on the three groups of the independent variables which affects the dependent variables (search engine ranking score, TrustRank, and link popularity).

Table 2 presents the results of two-way analysis of variance on the independent variables which affects search engine ranking score. Where the number of independent variables belong to table 1.

f = *F* test; which is used to compare the statistical models to find the best fit population from which the data were sampled.

Degrees of freedom (*df*); if there are *N* observations in total, *df*total = *N* – 1.

P-value; It is a value used to evaluate the statistical standards, when *P*-value< 0.05, we reject the null hypothesis [11].

Kappa statistic (*KS*); computes the percentage of error reduction compared to all errors in the classification sample. When there is no agreement between two raters *KS* is zero or close to zero, and when *KS* value is close to 1 this mean we have a perfect agreement between two raters [1].

Mean absolute error (*MAE*); measures the average of the errors in a set of the estimation, to show how much the estimation relatives to the actual outcomes [1].

Root Mean Squared Error (*RMSE*); measures the average of the errors, through measuring the difference between estimates and corresponding observed values. The range of *RMSE* from 0 to ∞; low values are better than high values [1].

TABLE II. Anova Results For Search Engine Ranking Score Model

Independent Variable Number	*df*	*f*	*P-value*	*KS*	*MAE*	*RMSE*	*Accuracy*
1-7, 9, 11, 13, 15, 17, and 18	12	278.11	0.006	0.92	0.03	0.19	0.96

According to table 2, we can find that *f* test is 278.1157 with *P*-value equals to 0.006751, which is less than significant value of *P*-value (0.05), so we reject the null hypothesis, and accept the alternative hypothesis which asserts the relationship between the fifteen content and link Web spam features and the search engine ranking score. This model yields an accuracy of 96%. Other performance measurements computed include; *KS*, *MAE*, and *RMSE* were close to optimal values.

Table 3 shows *ANOVA* results of the independent variables which affect TrustRank scores, with the same statistical measurements, that used before.

TABLE III. ANOVA RESULTS FOR TRUSTRANK MODEL

Independent Variable Number	*df*	*f*	*P-value*	*KS*	*MAE*	*RMSE*	*Accuracy*
1-7, 9-12, 13, 15, 17, and 18	14	5467.138	0.000	0.99	0.007	0.037	0.99

Table 3 shows that *f* test is equal to 5467.138 with *P*-value is equal to 0.000, which leads to the rejection of the null hypothesis, and accept the alternative hypothesis. The TrustRank model yields an accuracy of 99%.

Table 3 shows improvement results of *KS* and other performance measurements than the results of table 2, which are very close to the optimal values. Table 4 shows the *ANOVA* results for the link popularity model.

TABLE IV. ANOVA RESULTS FOR THE LINK POPULARITY MODEL

Independent Variable Number	*df*	*f*	*P-value*	*KS*	*MAE*	*RMSE*	*Accuracy*
2, 4, 6, 8, 10, 12, 14, and 16	7	2205.75	0.000	0.77	0.14	0.28	0.88

According to Table 4, *f* test is equal to 2205.754 with *P*-value is equal to0.000, so the null hypothesis has to be rejected. The link popularity model yields an accuracy of 88%. *KS* yields an accuracy of 0.77 which considered a significant value, since it is close to 1.

The performance measurement; *MAE* and *RMSE* yields accepted values (close to zero). The comparison of the results of the three previous tables reveal that TrustRank and search engine ranking score models yields nearly the same results, followed by the link popularity model.

V. CONCLUSION

Website masters and developers struggle to improve their Websites' visibility; such actions help to increase the value of the search engine ranking score, trust rank, and link popularity and give them better opportunities in e-commerce marketing and advertisement campaigns.

In this paper, we proposed an Online Arabic Web Spam Detection System (*OLAWSDS*), which uses features extracted from content and links and benefits from client/server models. SPSS package is used to evaluate our proposed system using the test of the analysis of variance (*ANOVA*). Results showed improved results in comparison with other approaches in terms of prediction accuracy or performance.

REFERENCES

[1] H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "A link and Content Hybrid Approach for Arabic Web Spam Detection, "International Journal of Intelligent Systems and Applications (IJISA), vol. 5, no. 1, pp. 30-43, 2013.

[2] W. Wang, G. Zeng, D. Tang, "Using evidence based content trust model for spam detection, "Expert Systems with Applications, vol. 37, pp. 5599-5606, 2010.

[3] H. Wang, Y. Lia, K. Guo, "Countering Web Spam of Link-based Ranking Based on Link Analysis," Procedia Engineering, vol. 23, pp. 310–315, 2011.

[4] C. Castillo, C. Corsi, D. Donato, "Query-log mining for detecting spam," Proceedings of the 4th international workshop on Adversarial information retrieval on the Web Pages AIRWeb '08, ACM, pp. 17-20, 2008.

[5] J. Martinez-Romo, L. Araujo, "Web spam Identification Through Language Model Analysis," Fifth International Workshop on Adversarial Information Retrieval on the Web AIRWeb '09, Madrid, Spain, pp. 21-28, 2009.

[6] J. Lin, "Detection of cloaked Web spam by using tag-based methods," Expert Systems with Applications, vol. 36, pp. 7493-7499, 2009.

[7] L. Araujo, J. Martinez-Romo, "Web spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models," IEEE Transactions on Information Forensics and Security, vol. 5, pp. 581-590, 2010.

[8] J. Lin, "Detection of cloaked web spam by using tag-based methods," Expert Systems with Applications, vol. 36, pp. 7493–7499, 2009.

[9] X. Liu, Y. Wang, S. Zhu, H. Lin, "Combating Web spam through trust-distrust propagation with confidence," Pattern Recognition Letters, vol. 34, no. 13, pp. 1462-1469, 2013.

[10] SPSS software, Retrieved October, 18, 2013, fromhttp://www-01.ibm.com/software/analytics/spss/

[11] R. N. Cardinal, "Graduate-level statistics for psychology and neuroscience ANOVA in practice, and complex ANOVA designs," 2004, fromhttp://egret.psychol.cam.ac.uk/psychology/graduate/Guide_to_ANOVA.pdf

[12] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," VLDB '04 Proceedings of the Thirtieth international conference on Very large data bases, vol. 30, pp. 576-587, 2004

AUTHORS PROFILE

Mohammed Naji Al-Kabi obtained his Ph.D. degree in Mathematics from the University of Lodz/Poland (2001), his masters degree in Computer Science from the University of Baghdad/Iraq (1989), and his bachelor degree in statistics from the University of Baghdad/Iraq (1981). Mohammed Naji AL-Kabi is an assistant Professor in the Faculty of Sciences and IT, at Zarqa University. Prior to joining Zarqa University, he worked many years at Yarmouk University in Jordan, Nahrain University and Mustanserya University in Iraq. He also worked as a part time lecturer at Jordan University of Science and Technology, Princess Sumaya University for Technology, and Sunderland university. AL-Kabi's research interests include Information Retrieval, Web search engines, Data Mining, Social media, Natural Language Processing and Software Engineering. He is the author of more than 66 peer reviewed articles in these topics. His teaching interests focus on information retrieval, Web programming, data mining, DBMS (ORACLE & MS Access).

Heider Wahsheh, born in Jordan, in August 1987, he obtained his Master degree in Computer Information Systems (CIS) from Yarmouk University, Jordan, 2012. Since 2013 Mr. Wahsheh starts working as a lecturer in the college of Computer Science at King Khalid University, Saudi Arabia. His research interests include: Information Retrieval, Data Mining, and Mobile Agent Systems.

Izzat Alsmadi. An associate professor in software engineering. Born in Jordan 1972, Izzat Alsmadi has his master and PhD in software engineering from North Dakota State University (NDSU), Fargo, USA in the years 2006 and 2008 respectively. His main areas of research include: software engineering, testing, metrics, and information retrieval.