EVALUATION OF WEB SPAM BEHAVIOUR ON ARABIC WEBSITES POPULARITY

Heider A. Wahsheh[1], Izzat M. Alsmadi[2], and Mohammed N. Al-Kabi[3]

[1-2]Computer Information Systems Department, IT & CS Faculty, Yarmouk University
Irbid /Jordan
[3] Faculty of Sciences & IT, Zarqa University
Zarqa /Jordan
[1]*heiderwahsheh@yahoo.com*
[2]*ialsmadi@yu.edu.jo*
[3]*malkabi@zu.edu.jo*

Abstract

The expansion of the information in the Web increased the Websites roles which present the importance of the information they provide. Many Webmasters try to inject misleading or marketing information and links that may eventually impact the value and reliability of provided information. Search engines and Web users need to detect and distract reliable information from such noise or extra information. In this paper, we collected the top 100 popular Arabic Websites from the Search Engine Results Page (SERP), using the top 10 most popular Arabic keywords, and used them as a case study. We evaluated these Websites using the main Web spam features. The results showed that some of these popular Websites are using illegal and Web spam techniques in order to boost their ranks within (SERP).

Keywords - Web spam features, Arabic Web spam, Website popularity.

1 INTRODUCTION

Websites are getting more and important and play major roles in public or private sectors' entities. The Website is not only a surplus source of information to advertise announcements, post company portfolio. The Website can be the main interface between the business and its clients or users. If e-commerce services are in place, the Website is then a major revenue source. Communications between business, its clients, suppliers or service providers can be largely conducted through this Website or related applications. As such, the need for such Website to be highly trustable is a must. Business clients and service providers need to trust that business Website and trust all information it has.

The internet has a number of Websites which categorize as spam Websites. The Web spam activities include a wide range of acts such as financial fraud and distributing of Malware to disrupt computer operation , also these Web spam acts include Email spam, selling of fake products such as wrist watches, perfumes, gadgets, clothes, shoes, software, hardware, music, health and personal care products …etc.

Evaluating Website popularity in the Search Engine Results Page (SERP) however, is not a straightforward process. In order to be able to come up with this small sized information a large number of techniques and parameters should be investigated and assessed concerning this Website to be able to come up with such judgment for a significant level of confidence. These techniques enhance Websites popularity in the search engines, some of these techniques considered as a legal such as: Search Engine Optimization (SEO) techniques and others considered as illegal (spam) and may cause the Website to be banned from the search engine listings. For example, a non spam Website is a Website that is known and has a fixed or significant number of users that frequently visit it. A non spam Website is a

Website that is reliable in terms of the information it is providing. Such information should be current, correct and relevant. Relevancy is related to the user search or what is the user looking for [1]. For example, a Website that includes vague pure marketing information with keywords that are irrelevant and meant only to drive users through search engines is not trust or reliable. The links between the Website, its authors, Web pages and information is like a connected triangle or square. This means that strength in one aspect can lead to strength in the other aspects and vice versa [1].

In this paper, we collected the top 100 popular Arabic Websites from the SERP, using the top 10 most popular Arabic keywords, and evaluated the impact of the Web spam techniques on the Websites popularity. The rest of the paper is organized as the following: The next section presents the main Web spam techniques. Section three shows some of the related studies of the Web spam. Section four presents the framework. Section five shows the experiments and results. Finally paper is concluded with a small section of the conclusion.

2 WEB SPAM TECHNIQUES

Web spammers apply many techniques to boost the spam Websites in the top of the SERP, they used the following three main Web spam techniques; content, link, and cloaking Web spam [1].

2.1 Content Web spam

Content Web spam refers for the manipulation on the HTML structure of the Web pages, in order to boost the rank of the Web pages [1, 2]. Keyword or key stuffing is considered as a main method in the content Web spam, where the Web spammers stuffed and duplicate the Arabic, English, and symbols words and characters in the main HTML elements such as: <Meta> elements, in order to achieve better visibility in SERP [3, 4]. The spammers try to increase the number of used <Meta> elements in the Arabic Web pages, in order to stuff more number of words and characters. Using high number of images which is not related to the surrounded content is considered as another spam method in the Arabic Web pages to increase the rank of the spam Web page within the SERP [3, 4]. When the images used as hyperlinks, the spammers trick the users with the content of images, while when users click on the images they get a spam Web pages which contradict to the desired Web page. The spammers benefit from this spam behaviour which considered as one of the pay per click (PPC) marketing techniques; by attracting more users to click on these spam images. So the spammers gain more revenues [4].

2.2 Link Web spam

The PageRank algorithm is considered as a main ranking algorithm, so the details about how this algorithm work is considered as a secret. It assumes that there is a random Web surfer, and it is based on the forward links. The PageRank represents the probability of Web surfer to randomly visit a Web page [5]. Web spammers use the link Web spam, in order to violate the PageRank algorithm. Link spam presents the technique which includes adding many irrelevant links to a Webpage [4, 6]. Those are irrelevant to the mother page and even not related to each other. This method helps the Website to build unnatural link popularity and the search engine will most likely notice the link neighborhood. This is also called un-natural linking and has a goal to have more incoming links pointing to their domain unethically using link farm or link buying [4].

2.3 Cloaking Web spam

Cloaking Web spam is a spam technique which based on using two different versions for the same Web pages, the first version present the high quality of customized content and links features to send it to the Web crawler to gain the highest possible rank. While the second version present the low quality Web pages (spammed Web pages), and hides the content from the visitors, which received to the user and appearing in the user browser [7].

3 RELATED WORK

The field of the Web spam is a popular subject. However, in this small related work section we will only refer to few papers with the focus on evaluating spam in Arabic documents, or evaluating the correlation between spam and the popularity.

The authors of this study already have a number of publications in evaluating both the content and link Arabic Web spam. In those papers authors collected and built large datasets of Web pages in Arabic with spam or possible spam. Authors used rules of the various number of content and link based features that defined by search engines. They applied number of machine learning algorithms on these datasets. The results showed that the Decision Tree in the most cases is the best to detect and evaluate the Arabic Web spam [3, 4, 8, 9, 10, 11, 12, 13, and 14].

The study of [6] dedicated for combating the link-based Web spam through using two proposed groups of the temporal features. The first group called In-link Growth Rate (IGR) which defined the ratio of the increased number of internal links in Web pages. While the second group called In-link Death Rate (IDR), which defined the ratio of the number of broken internal links to the number of all internal links in the Web page. They used support vector machines (*SVM*) as a spam classification model, and the results achieved an accuracy of 40% – 60% in detecting link-based Web spam.

The study of [2] applied a language model approach which extract various of content and link features from the Web page, to provide high quality of detecting Web spam.This model applied Kullback-Leibler (KL) divergence on each two Web pages which linked together by a hyperlink, to determine the relationship between them. The experiments used two large public datasets; WEBSPAM-UK2006 and WEBSPAM-UK2007, and the results improvement the F-measure by 6% in WEBSPAM-UK2006, and near 2% in WEBSPAM-UK2007.

The study of [15] presented different categories for Web spam features based on recent advances in Web spam filtering. Three of machine learning algorithms (i.e. ensemble selection, LogitBoost and Random Forest) were used. The conducted tests were applied on the two well known available datasets WEBSPAM-UK2007 and the Discovery Challenge dataset DC2010. The tests used ensemble classifier to detect spammed Web pages, and the improvement results ranged between 5-7.5%.

In their study [16], the researchers proposed a framework for splog detection by monitoring the top-ranked results. The framework arranged the sequence of temporally queries and detected splogs based on the temporal behaviour. The experiments showed a high accuracy on splogs detection.

A new method proposed by [17] to detect the Web spam problem through two players game to identify the spam Web pages within search results. The novel game asks player to classify the Web pages as relevant, irrelevant, or passing to specific queries. This method was considered effective as truthfully voting Web spam algorithm.

The study of [18] dedicated for detecting cloaking techniques, through using three of proposed tags methods; TagDiff2, TagDiff3, and TagDiff4. The proposed methods responsible for find the differences in the HTML tags for both user and Web crawler versions of the Web page. The experiments results in precision and recall showed that the tag methods exceed the content and link detecting methods

4 FRAMEWORK

This study provides an evaluation of the Web spam behaviour on the Arabic Websites popularity. We analyzed the three main elements (Web search engines, Web users, and Webmasters) that affected on the Websites popularity on the SERP. Figure 1 presents the interaction of these elements.

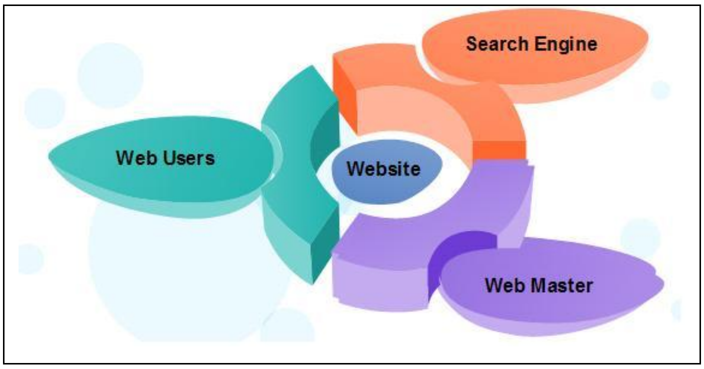


Fig. 1. The interaction of the Popularity Website Elements.

Each element presents in Figure 1 includes set of properties which considered to increase the Websites popularity.

Our framework includes the following steps:

1. Collecting the top 100 Arabic popular Websites from the SERP, using the top 10 most popular Arabic keywords from the SERP [14].
2. Extracting a set of content and link Web spam features using the proposed tool in [4].
3. Evaluating the Web spam behaviour on the Arabic Websites popularity.

4.1 Case Study

In this paper we extract from each Website a 10 Web pages in the average, including the homepage, to ensure that if the spam behaviour covers the whole of Websites or not. So we used 2,000 Arabic Web pages as a case study.

We removed from our case study the Websites (with .edu domain name) which considered as a non spam Web pages, and these Web pages considered outside this study.

4.2 Web spam features

In their study [4], the authors proposed a system which can extract the set of content and links Arabic Web spam features. The Arabic Web spam content and link metrics can be analyzed in different behaviour of the Website popularity

Table 1 shows the behaviour of some Arabic content and link metrics with the Website popularity.

Table 1. The behaviour of some Arabic content and link features with the Website popularity

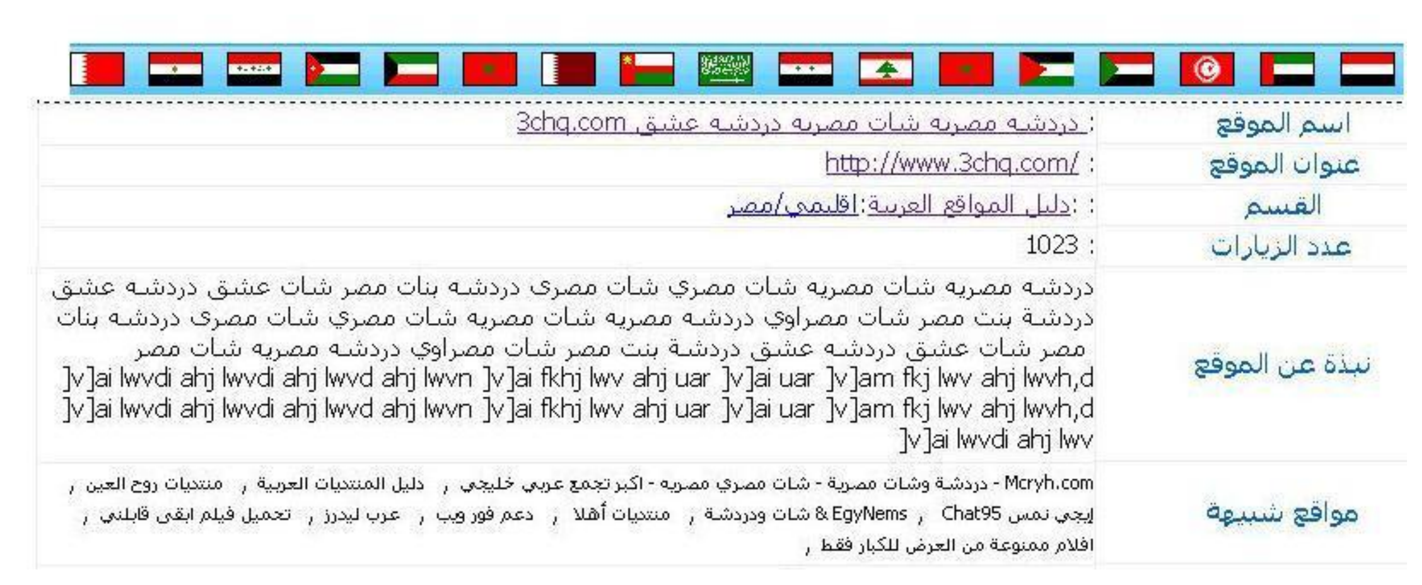
Arabic Web spam features	Search Engine	Web user	Webmaster
Meaningless (Arabic/English) keyword stuffing and normal keyword stuffing.	Mislead the search engine, and damage the quality of Web pages in SERP.	-	Act as spammers.
Size of compression ratio for Web pages.	Increasing the compression rate in a Web page used to hide the redundant content.	Attractive users to access these Web pages.	Used as one of SEO tips to attract the users to visit the Web page.

Size of hidden text.	Mislead the search engine.	-	Act as spammers.
Number of Images or images links.	Mislead the search engine.	Attractive users to access these Web pages.	Act as spammers.
Average length of Arabic/English words	Increase the lengths will increase the weights of the words in SERP.	-	Used as one of SEO tips to attract the users to visit the Web page.
URL length.	Increase the URL length can indicate that it is used spam words.	Short URL length, easy to remember.	Act as spammers. Increased the URL length.
External links	Small number of external links indicates to trust (good) Web pages (within Web spam considerations).	-	Need to reduce the number of external links.
Internal links	High number of internal links indicates to trust (good) Web pages (within Web spam considerations).	-	Need to increase the number of internal links (within Web spam considerations).

5 EXPERIMENTS AND RESULTS

In this study we analyzed and extracted the Arabic Web spam features, to evaluate the spam behaviour on our case study. We found that 37 Websites from our case study used the Arabic Web spam features to boost their ranks, so the Arabic Web spam features have a major affect on increasing the Arabic Websites popularity.

Figure 2 shows a portion of an example of Arabic spammed Web page, with high popularity for the 'chat' (شات ، دردشة) word. This Web page available online[1], and presents the description of another Website[2].



:دردشه مصريه شات مصريه دردشه عشق 3chq.com	اسم الموقع
http://www.3chq.com/ :	عنوان الموقع
: :دليل المواقع العربية:اقليمي/مصر	القسم
1023 :	عدد الزيارات
دردشه مصريه شات مصريه شات مصري شات مصرى دردشه بنات مصر شات عشق دردشه عشق دردشة بنت مصر شات مصراوي دردشه مصريه شات مصريه شات مصري شات مصرى دردشه بنات مصر شات عشق دردشه عشق دردشة بنت مصر شات مصراوي دردشه مصريه شات مصر]v]ai lwvdi ahj lwvdi ahj lwvd ahj lwvn]v]ai fkhj lwv ahj uar]v]ai uar]v]am fkj lwv ahj lwvh,d]v]ai lwvdi ahj lwvdi ahj lwvd ahj lwvn]v]ai fkhj lwv ahj uar]v]ai uar]v]am fkj lwv ahj lwvh,d]v]ai lwvdi ahj lwv	نبذة عن الموقع
Mcryh.com - دردشة وشات مصرية - شات مصري مصريه - اكبر تجمع عربي خليجي , دليل المنتديات العربية , منتديات روح العين , إيجي نمس Chat95 , EgyNems & شات ودردشة , منتديات أهلا , دعم فور ويب , عرب ليدرز , تحميل فيلم ابقى قابلني , افلام ممنوعة من العرض للكبار فقط ,	مواقع شبيهة

[1] http://dir.mobi4all.net/show29858.html

[2] http://www.3chq.com

Fig. 2. An example of spam Web page with 'chat' (شات، دردشة), word.

Figure 2 presents an example of spam Web page, which used the meaningless (Arabic/English) keyword stuffing and normal keyword stuffing for the 'chat' (شات ، دردشة) word in the Web pages.

Table 2 summarized the other content and link Arabic web spam features for our example of the spammed Web page.

Table 2. Arabic Web spam behaviour for our example

Arabic Web spam feature	The values of Arabic Web spam features
Size of the compression ratio.	83% of the Web page size
Size of hidden text.	0
Number of images or images links.	27
Average length of Arabic/English words.	6.48
URL length (characters).	31
Meta words.	60
Total words in the Web page.	311
Duplicate words in the Web page.	261
Total words in the title.	15
External links.	8
Internal links.	81

Table 2 shows that our example applied the Arabic Web spam behaviour for the Arabic content Web spam features. Although that the values of the external and internal links appeared in the non spam behaviour, we tracked these links, and especially the links that appearing in the figure 2 as a 'Similar sites' (مواقع شبيهة), and we found that this Web pages linked with the other Web pages, and composed the link farm of the spammed Web pages (i.e. the Website (http://www.3chq.com) presents the spam behaviour in the title, with 11 words, in order to increase it's rank).

6 CONCLUSION

The spam Websites have a negative impact on the Internet users, and lead to a decrease in the public confidence, and a decrease in the productivity and safety of the search engines. These Websites usually used for commercial advertisements and financial goals, so they try to optimize their popularity and visibility on the SERP. In this study we collected and analyzed the top 100 Arabic Websites from the SERP. We evaluated the Web spam behaviour on the popularity of these Websites.

The results showed that some of those Arabic Websites are using illegal techniques and Web spam features in order to boost their ranks within SERP. We noticed however that the majority of the popular Websites in this region falls under the category of entertainment and social Websites. In comparison to the official Websites for e-government portals, universities, banks, etc, such commercial and financial Websites are known generally to care less on following SEO and search engines' guidelines. This may explain that such spam usage may fall under non intentional or negligent acts.

7 REFERENCES

[1] Z. Gyongyi, and H. Garcia-Molina, " Web spam taxonomy ". In Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web, Chiba, Japan. pp 1-9, 2005.

[2] J. Martinez-Romo, and 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.

[3] H. Wahsheh, I. Abu Dosh, M. Al-Kabi, I. Alsmadi, and E. Al-Shawakfa, "Using Machine Learning Algorithms to Detect Content-based Arabic Web spam" Journal of Information Assurance and Security, vol. 7 pp. 14-24, 2012.

[4] 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.

[5] M. Selvan, A. Sekar, and A. Dharshini, "Survey on Web Page Ranking Algorithms" International Journal of Computer Applications, vol. 41, no. 19, pp 1-7, 2012.

[6] G. Shen, B. Gao, T. Liu, G. Feng, S. Song, and H. Li, "Detecting Link spam using Temporal Information". In Proceedings of the Sixth International Conference on Data Mining Pages ICDM '06, IEEE, pp 1049-1053, 2006.

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

[8] H. A. Wahsheh, and M. N. Al-Kabi, "Detecting Arabic Web spam". The 5th International Conference on Information Technology, ICIT 2011, Amman-Jordan, pp 1-8, 2011.

[9] R. Jaramh, T. Saleh, S. Khattab, and I. Farag, "Detecting Arabic spam Web pages using Content Analysis" International Journal of Reviews in Computing, vol. 6, pp 1-8, 2011.

[10] M. Al-Kabi, H. Wahsheh, A. AlEroud, and I. Alsmadi, "Combating Arabic Web spam Using Content Analysis". 2011 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman Jordan, pp 1-4, 2011.

[11] H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "Spam Detection Methods for Arabic Web Pages". First Taibah University International Conference on Computing and Information Technology ICCIT, Al-Madinah Al-Munawwarah, Saudi Arabia, vol. 2, pp 486-490, 2012.

[12] H. Wahsheh, M. Al-Kabi, and I, Alsmadi, "Evaluating Arabic spam Classifiers Using Link Analysis". In Proceeding of the 3rd International Conference on Information and Communication Systems ICICS'12, ACM, pp 1-5, 2012.

[13] M. Al-Kabi, H. Wahsheh, I. Alsmadi, E. Al-Shawakfa, A. Wahbeh, and A. Al-Hmoud, "Content Based Analysis to Detect Arabic Web spam" Journal of Information Science, vol. 38, pp 284-296, 2012.

[14] H. Wahsheh, I. Alsmadi, and M. Al-Kabi, "Analyzing the Popular Words to Evaluate spam in Arabic Web Pages" IJJ: The Research Bulletin of JORDAN ACM – ISWSA, vol. 2, no. 2, pp 22-26, 2012.

[15] M. Erdelyi, and A. Benczur, " Temporal Analysis for Web spam Detection: An Overview". In proceedings of the 1st Intl. Temporal Web Analytics Workshop TWAW 2011, Hyderabad, India. pp 17-24, 2011.

[16] L. Zhu, A. Sun, and B. Choi, "Detecting spam blogs from blog search results" Information Processing and Management: an International Journal, vol. 47, no. 2, pp 246-262, 2011.

[17] M. Goodstein, V. Vassilevska, "A Two Player Game To Combat Web spam" School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, 2007.

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