King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Evaluation of Spam Impact on Arabic Websites Popularity

Mohammed N. Al-Kabi [a], **Izzat M. Alsmadi** [b,*], **Heider A. Wahsheh** [c]

[a] *Faculty of Sciences and IT, Zarqa University, Zarqa, Jordan*
[b] *Computer Science Department, Boise State University, Boise, ID 83725, USA*
[c] *Computer Science Department, College of Computer Science, King Khalid University, Abha, Saudi Arabia*

Received 23 July 2013; revised 5 March 2014; accepted 15 April 2014
Available online 6 April 2015

KEYWORDS

Web metrics;
Web spam;
Link spam;
Arabic Web spam;
In-link;
Out-link

Abstract The expansion of the Web and its information in all aspects of life raises the concern of how to trust information published on the Web especially in cases where publisher may not be known. Websites strive to be more popular and make themselves visible to search engines and eventually to users. Website popularity can be measured using several metrics such as the Web traffic (e.g. Website: visitors' number and visited page number). A link or page popularity refers to the total number of hyperlinks referring to a certain Web page. In this study, several top ranked Arabic Websites are selected for evaluating possible Web spam behavior. Websites use spam techniques to boost their ranks within Search Engine Results Page (SERP). Results of this study showed that some of these popular Websites are using techniques that are considered spam techniques according to Search Engine Optimization guidelines.

1. Introduction

Websites strive to be popular and make themselves visible to Web search engines. Internet visibility depends on Website traffic. Traffic is determined by the number of users or visitors for a particular Website. Search engines work as mediators between users and Websites. Most of Web users use the search engines as guiding tools to the relevant Web documents based on their information needs. Search engine users have to formulate queries expressing their information needs and submit these queries to search engines to retrieve Search Engine Results Page (SERP). There are several techniques that can be used to enhance Website visibility to search engines. Some of these techniques are legal and recommended by search engines and known as Search Engine Optimization (SEO) recommendations. Others are considered illegal and may cause the Website that uses them to be banned from the listings of any search engine when discovered such spam behavior. For

[*] Corresponding author.
E-mail addresses: malkabi@zu.edu.jo (M.N. Al-Kabi), izzatalsmadi@boisestate.edu (I.M. Alsmadi), heiderwahsheh@yahoo.com (H.A. Wahsheh).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.04.005
1319-1578 © 2015 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

example, Google presents a number of beneficial guidelines showing how a Webmaster or an administrator can raise legally the rank of their Web pages.

In Web or link spam, a Website or a Web page is injected with irrelevant content to raise falsely its popularity. Real Website popularity should come from real users who are visiting the Website or real Websites which are pointing to or linking to other related Websites. Non spam Websites usually refer to other non-spam Websites if the target Websites contain additional useful information or provide additional services to its visitors. Using spam techniques within Web pages may lead temporarily to raise their ranks. Eventually both users and search engines find out that spam Website is misleading them and may eventually hurt search engine credibility or reputation, besides hurting the credibility of these spam Websites. Fake traffic, which is based on unrealistic artificial traffic, can be used to deceive search engines which consider the popularity as one of the important parameters in the ranking of their results. Such act may eventually hurt the popularity and credibility of those Websites. In general, defining spam and rules for spamming facilitate the spam identification by Web search engines. For example, Google defines the following practices to be spam techniques (Gyongyi and Garcia-Molina, 2005):

- Hidden texts or links.
- Cloaking or tricky redirects.
- Automated queries to the search engine.
- Pages loaded with irrelevant keywords.
- Multiple pages, sub-domains, or domains with substantially duplicate content.
- “Doorway” pages created particularly for search engines. These are pages which have been designed to rank high on search engines. They are then set to redirect visitors to the actual Website.

The main challenge in the research related to Web spam techniques can be summarized by the ambiguity of the rules used by Web search engines to identify spam Web pages. This is so because these rules are considered by search engines as part of their ranking algorithms, and therefore they are classified and not publically exposed.

There are also other related issues or challenges such as facing a contradiction between spamming techniques and SEO optimization guidelines. Moreover, the adopted spam rules used by different Web search engines to identify spam Web pages are different, and not unified. Therefore, a certain Web page maybe considered by a certain search engine as a spam while it is ranked within the top 10 SERP for another search engine.

The term “Spamdexing” is used to describe techniques used to artificially raise the perceived relevancy of inferior Websites (Gyongyi and Garcia-Molina, 2005).

In this paper, we evaluate the level of using spam techniques in most popular Arabic Websites (listed according to Alexa.com for ranking Website popularity). Top Websites according to Alexa.com are evaluated according to several guidelines against conducting spam techniques or behaviors.

The rest of the paper is divided as follows: Section 2 presents selected related works on Web spam detection studies. Section 3 discusses spam techniques with the main ranking algorithms. Section 4 presents experiments and results. Section 5 presents the conclusion of this paper.

2. Related Work

The literature includes several research publications related to the subject of Web spam where this topic is studied from different perspectives. This Section presents few of these studies which are closely related to the paper subject: Web spam detection, to detect both Arabic and non-Arabic Web spam, and those studies dedicated to the evaluation of the correlation between spam and popularity.

There are several publications related to detection of Arabic content and link based Web spam conducted by this paper authors. The study of Wahsheh et al. (2013a) used the dataset of top 100 popular Arabic Websites from the search engine results pages, which were collected based on the popular Arabic key words. The evaluation of these Websites is conducted by extracting the main Web spam features of Wahsheh et al. study (Wahsheh et al., 2013b) through three main Websites’ elements (Web users, search engines and Web masters). The study of Wahsheh et al. (2013b) proposed an Arabic content/link Web spam detection system, which extracts proposed Arabic Web spam features, and adopts three classification techniques and machine learning algorithms to identify spammed/non-spammed Arabic Web pages. Results showed also that while there are some common behaviors among all languages for spam, however, each language, particularly Arabic, may have unique rules that can be used or abused by spammers (Wahsheh et al., 2013b). There are also other studies that are related to the use of spamming within certain Arab nation, such as the study of Al-Kadhi (Al-Kadhi, 2011). In his study he conducted a comprehensive survey study to determine the state of the use of spamming in the Kingdom of Saudi Arabia (KSA). His study includes all related statistics to spam and refers to the measurements of specialized companies to the percentages of spamming behaviors in KSA.

One of the main goals of link and content Web spam is to enhance the popularity of the Web pages which adopt them. In order to limit the effect of these techniques the paper of Schwarz and Morris (2011) proposed the augmentation of search results with additional features in order to make the results more accurate and thus to reduce the effect of spam techniques on SERP. Their study aims to help users and visualization techniques to measure the credibility of Websites. Website credibility measures several aspects related to the level of trust that users can have on Websites. Both credibility and popularity measure how many users are visiting the subject Website and how many other Websites are pointing to it.

The study of Bhushan and Kumar (2010) also discussed the issue of Website ranking, credibility and some of the factors that may have a positive impact on ranking. The studies of Moe (2011) and Li and Walejko (2008) discussed the issue of spam in Weblogs and their ability to bias or produce incorrect or inaccurate results. The study of Goodstein and Vassilevska (2007) proposed a new truthfully voting algorithm for Web spam detection through a 2-player game, where each player has to classify the Web pages as relevant, irrelevant, or passing to specific queries. Another study based on the feedback of the users that is converted to the query log is conducted by Castillo et al. (2008). For each user, a query log file was assigned. Researchers in the paper applied two approaches: Web spam detection and query spam detection.

The study of Shen et al. (2006) studied the link-based Web spam through using the link-based temporal information. Temporal features are used to detect the spam behavior. These features are divided into two groups; the first one is called Internal link Growth Rate (IGR) which shows the ratio of the increased number of internal links in Web pages, and the second one is called Internal link Death Rate (IDR) which defines the ratio of the number of broken internal links to the number original internal links in the Web pages. The experimental tests used the support vector machines (SVM) classifier to evaluate the proposed approach and achieved a relatively high accuracy percentage (40–60%).

3. Spam Techniques with Ranking Algorithms

Spammers use various spamming techniques (i.e. hiding links, cloaking, link farming, and keyword stuffing) to deceive search engines and increase their Website ranks.

These spamming techniques succeed in many cases to deceive the ranking algorithms adopted by different search engines. The success of spamming techniques to deceive a search engine yields non-relevant results to the query, and this damages the reputation of search engine.

This Section presents three important ranking algorithms (Term Frequency-Inverse Document Frequency, PageRank, and Hyperlink-Induced Topic Search), and shows how spammers attempt to deceive these three algorithms to gain the best possible rank for the spammed Web pages in the SERP.

3.1. The Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (*TF-IDF*) is a numerical statistic weight used to evaluate the importance of a word in a certain document or in a collection of documents.

The study of Baeza-Yates and Ribeiro-Neto (2010) presents four formulae for Term weighting; F_i, *TF*, *IDF*, and *TF-IDF* as shown in the following mathematical equations:

Let,

k_i be an index term and d_j is a document.
$V = \{k_1, k_2, \ldots, k_t\}$ be the set of all index terms.
($w_{i,j} \geqslant 0$) be the weight associated with (k_i, d_j).

The weights $w_{i,j}$ are computed using the frequencies of occurrence of the terms within documents. $f_{i,j}$ is the frequency of occurrence of index term k_i in the document d_j. So the total frequency of occurrence F_i of term k_i in the collection is defined as shown in formula (1):

$$F_{i,j} = \sum_{j=1}^{N} f_{i,j} \tag{1}$$

where N is the number of documents in the collection.

The study of Baeza-Yates and Ribeiro-Neto (2010) presents the Luhn assumption which indicates that the weight of $w_{i,j}$ of index term k_i that occurs in the document d_j is relative to the Term Frequency $f_{i,j}$. This assumption means that increasing an occurrence of the term in the document, leads to get the highest weight.

The formula of Term Frequency *TF* is presented in formula (2):

$$TF_{i,j} = f_{i,j} \tag{2}$$

while the variant of *TF* weight is presented in formula (3):

$$TF_{i,j} = \begin{cases} 1 + \log(f_{i,j}) & \text{if}\,(f_{i,j} > 0) \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

The formula of Inverse Document Frequency (IDF) is presented in formula (4):

$$IDF_i = \log \frac{N}{n_i} \tag{4}$$

where *IDF is the i* inverse document frequency of term k_i.

The best known term weighting schemes use combination weights of $TF_{i,j}$ and IDF_i factors.

The Term Frequency-Inverse Document Frequency (*TF-IDF*) formula is shown in the following formula (5):

$$w_{i,j} = \begin{cases} (1 + \log(f_{i,j})) \times \log_2(\frac{N}{n_i}) & \textit{if}(f_{i,j} > 0) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $w_{i,j}$ is the term weight of the term k_{i} in the document d_j which refers to (*TF-IDF*) weighting scheme; $f_{i,j}$ is the frequency of occurrence of index term k_i in the document d_j (Baeza-Yates and Ribeiro-Neto, 2010).

Spammers try to increase the *TF-IDF* scores in their spam content-based Web pages. They used the following techniques:

3.1.1. Hiding links, texts and tags.

The goal of this technique is to deceive the search engines to refer to URLs that are not visible to normal users. This can be done through embedding them in very small pictures for example. When text is hidden off page or it uses the same color as the page background, search engines consider it spam (Gyongyi and Garcia-Molina, 2005).

3.1.2. Keyword stuffing

Spammers use many repeated and unrelated words in tags of an HTML such as: the <body> tag, Anchor text, URL, Headers (<h1> … <h6> tags), <meta> tags, and the Web page <title>, with many repeated and unrelated words in order to gain a higher TF-IDF score (Gyongyi and Garcia-Molina, 2005).

3.2. Hyperlink-Induced Topic Search (HITS) Algorithm

Hyperlink-Induced Topic Search (HITS) algorithm, is a well-known method to find the Hubs and Authoritative Webpages, that, is introduced by Jon Kleinberg in 1999, as a link analysis algorithm. It is proposed before the PageRank algorithm used for ranking Web pages (Selvan et al., 2012). HITS divided the Web pages into two main types: the first one is called hubs; which indicates the Web pages that work as large directories, that do not actually hold the information. Rather it points to many authoritative Web pages, which actually hold the information. So a good hub represented a Web page that points to many other Web pages. The second type is called authority Web page which holds the actual information, and a good authority is represented as a Web page which was pointed to by several hubs (Selvan et al., 2012; Jayanthi and Sasikala, 2011).

HITS compute two values for each Web page: the first value is for the authority which represents the score of the content-based Web page, and the second value is for the hub, which estimates the score of its links to other Web pages (Selvan et al., 2012).

Formula (6) presents the Authority Update Rule:

$\forall p$, we compute $A(p)$ to be:

$$A(p) = \sum_{i=1}^{n} H(i) \tag{6}$$

where $A(p)$ is the Authority for p Web page; n is the total number of Web pages that are linked to p; I is the Web page linked to p; and the $H(i)$ is the hub value for the I Web page that points to p (Selvan et al., 2012).

Formula (7) expresses the Hub Update Rule as shown below:

$\forall p$, we compute $H(p)$ to be:

$$H(p) = \sum_{i=1}^{n} A(i) \tag{7}$$

where $H(p)$ is the Hub for p Web page; n is the total number of Web pages p connected to; I is a page which p connects to; and the $A(i)$ is the Authority values for I page (Selvan et al., 2012).

The Web page is classified as a good hub if it points to many good authoritative, and the Web page is classified as a good authority if it is referred to by many good hubs. The hub values can be spammed through the link spam farms by adding the spam outgoing links to the reputable Web pages. So that spammers attempt to increase the hub values, and attract several incoming links from the spammed hubs to point to the target spam Web pages (Gyongyi and Garcia-Molina, 2005).

3.3. PageRank Algorithm

PageRank was proposed and developed by Google's founders (Larry Page and Sergey Brin) as a part of a research project about a new kind of search engines. It defines a numeric score which measures the degree of Web pages relevance to particular queries. It is important due to the high score value of PageRank that determines the list of SEPR for corresponding queries (Kerchove et al., 2008).

PageRank can be seen as a model of user behavior. It assumes that there is a random Web surfer, starts from randomly Web page. Web surfers usually keep clicking on the forward links, and when the time passes they get bored and choose another random Web page. Therefore, the PageRank score represents the probability of Web surfer to randomly visit a Web page (Kang et al., 2011).

The PageRank algorithm is considered as one of the main successful factors in Google. So this algorithm and how it works is considered as a top secret. The last revealed algorithm from Google indicates that the PageRank algorithm is a link ranking one, which takes the number of internal links as an important factor in page popularity. PageRank gives each page a score that determines the popularity of that page. The overall score of a page p is determined by the importance (PageRank scores) of pages which have out links to that page p (Kang et al., 2011). The generic formula which appears in the literature for calculating PageRank score for a page p is shown in the following equation:

$$r(p) = \alpha \times \sum_{(q,p)} \frac{r(q)}{w(q)} + (1 - \alpha) \times \frac{1}{N} \tag{8}$$

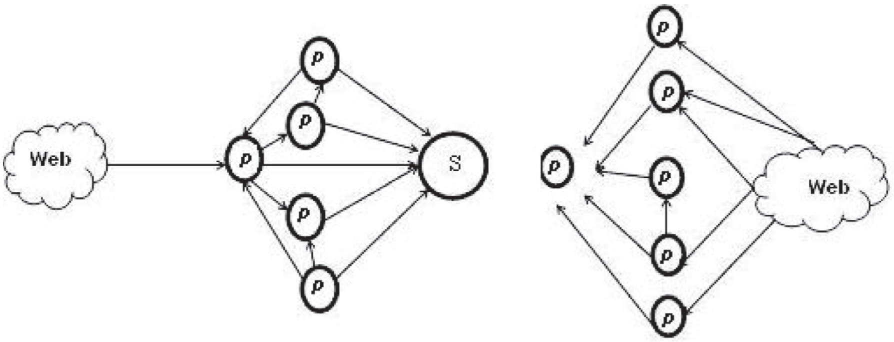
where $r(p)$ is the PageRank value for a Web page p; $w(q)$ is the number of forward links on the page q; $r(p)$ is the PageRank of page q; N is the total number of Web pages in the Web; α is the damping factor; (q, p) means that Web page q points to Web page p (Berlt et al., 2010).

A Web page with a high PageRank score will appear at the top of the list of SEPR as a response to a particular query. Despite this success for those search engines that use PageRank as a ranking algorithm, spammers and malicious Web masters use some of PageRank algorithm problems to boost the rank of their Web pages illegally by using techniques that violate the SEO tips, in order to gain more visits from Web surfers to their Website. Since PageRank is based on the link structure of the Web, it is therefore useful to understand how addition or deletion of hyperlinks influences it.

The degree of success in the link structure modifications is based on the degree of Web page accessibility by spammers. In most cases, the Web pages cannot be modified by spammer, so it is difficult for spammers to modify the link structures for such Web pages. Some Web pages on the other hand are partly accessible by spammers, hence, in a limited way spammers can post comments on such Web pages, such comments may carry an external link from blog site to their spam page (Gyongyi and Garcia-Molina, 2005). The third kind of Web pages to which spammers have full access is those Web pages owned by spammers. In such Web pages spammers try to create a link structure that works as a spam link farm, which is defined in Du et al. (2007) as a heavily connected Web page, created intentionally with the purpose of tricking a link-based ranking algorithm. In such case spammers will create a link structure that consists of few boosting Web pages that may refer directly to each other and to the spam pages in order to achieve some advantages by search engine ranking algorithms. In the study of Du et al. (2007) it is shown that spammers can build different structures for a spam farm, and such a farm structure may be changed periodically in terms of the number of internal and external links, that is when spam filters drop spam links it is expected from the spammers to change their link structure by adding new links to their spam farm structure.

Fig. 1 shows a sample Web graph with two structures, the one on the left presents a set of densely connected Web pages (p), where each one has links to another as well to a spam page which is the target whose rank is to be boosted. It appears in Fig. 1a (left), which has few links to the rest of the Web, and its goal is to boost the rank of spam Web pages by having too many internal links for its boosting neighbor's Web pages. On the other hand, Fig. 1b (right) has a normal structure and consists of a set of Web pages which have enough connections with the rest of Web graph. The differences between these two structures attract researchers to study the properties of these two structures and the variations of the structure appear in the left Web graph (Du et al., 2007).

It is known from the previous discussion that spammers have partial accessibility to some external Web pages that may have a good ranking score in search engine ranking vector. So it is expected from spammers to post links to those Web pages, because having a huge number of internal links on their spam page may achieve some improvement on its rank.



(a) Spam link farms structure (left) (b) Normal link structure (right)

Figure 1 Two main Web graph structures (Du et al., 2007).

Fig. 2 exhibits an example of a Web graph in which spammers make an attempt to boost the rank of spam page (*S*). The link structure used in Fig. 2 is an example of optimal link spam farm used in Gyongyi and Garcia-Molina (2005), Largillier and Peyronnet (2011) in which the authors proved how spammers can achieve benefit of having this structure. The structure consists of one target spam Web page (*S*). The spammers' goal is to boost the PageRank of this target Web page by pointing to page *S* using a set of Web pages $X = \{x_1, x_2, x_3\}$ in which the spammers have some accessibility (i.e. posting comments, adding links), spammers have also a full access to Web pages owned and created by them. So, the spammers also use their own set of Web pages $Y = \{y_1, y_2\}$. This set of Web pages is used mainly to post links to the target page *S* in order to boost its rank. Spammers will also add some external links from page *S* to the Web pages: $Y = \{y_1, y_2\}$, however no out links will be posted on Web pages $Y = \{y_1, y_2\}$, except those to the target page *S*.

The total PageRank score of the page *S* is maximized by the set of accessible $(x_1 \ldots x_3)$. The score that the target Web page gains from the boosting Web pages is calculated using the formula (9):

$$\sum_{i=1}^{3} \frac{r(p)}{out(x_i)} \tag{9}$$

where $r(p)$ is the PageRank; and $Out(x)$ the number of accessible Web pages (Zhou and Pei, 2009).

Every accessible Web page linked to the target spam page may have some contribution to its PageRank score. Such links are called hijacked links (Du et al., 2007). The total of PageRank scores of popular Web pages that have links (hijacked links) pointing to target spam Web pages is called leakage. The leakage gained by hijacked links is not known by spammers; however, their goal is to have as much hijacked links as it is possible.

The target page *S* PageRank score can be also maximized if that page points to all Web pages created and maintained by spammers (boosting Web pages), given that those Web pages have no internal links except those from the *S*. So the search engine will, reach the spam farm through one of its hijacked links. It is possible then to crawl boosting Web pages through the external link from the target spam page (Chung et al., 2010).

Finally, the *S* rank score can be also maximized if the set of owned Web pages $\{y_1, y_2\}$ has only external links to the target page *S*. This requires no links between boosting Web pages to each other. It requires also no hijacked links from outside world to the boosting Web pages (except from the *S*). The targeted page actually needs to point to all boosting Web pages to improve its PageRank score and to make every single Web page in the whole spam farm accessible by search engine crawler (Du et al., 2007).

4. Experiments and Results

The following three main steps summarize the experiments conducted in this study:

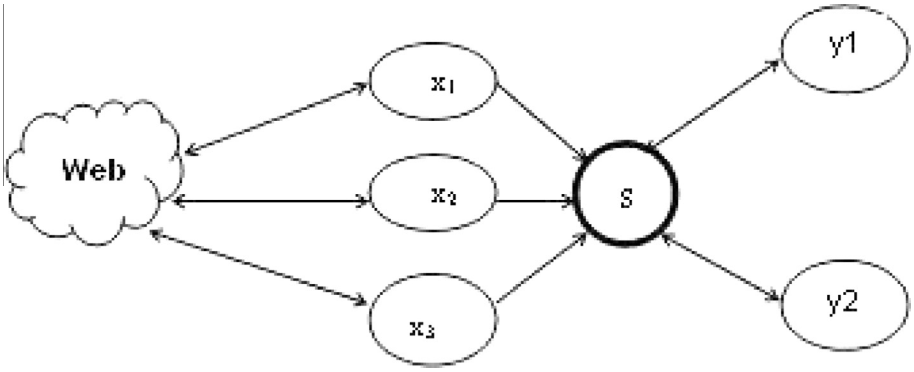


Figure 2 Optimal link spam farm structure (Gyongyi and Garcia-Molina, 2005).

1. Collect the most popular Arabic Websites and pages based on Alexa.com traffic and popularity ranking Website.
2. Analyze and extract the main Arabic content/link Web spam features from collected Websites, using the tool described previously in Wahsheh et al. (2013b).
3. Evaluate the collection of the most popular Arabic Web pages against the listed Arabic content/link Web spam features (Table 1).

During 2012 fourth quarter, we collected the dataset used in this study. This dataset has the top popular Arabic Websites according to Alexa.com ranking in that period. It should be noted however that such ranking list maybe frequently changed and updated which may change the rank of viewed pages or even change partially the list.

A previous study of the authors (Wahsheh et al., 2013b) proposed an Arabic content/link Web spam detection system, which consists of the following main parts:

1. An Embedded Web crawler, which is used to download the Web pages and parse all the Web pages elements (i.e. images, content, and links).
2. Arabic Web spam dataset, which contains 23,000 Arabic Web pages; 18,000 of them are used as a training dataset, while the rest are used as the testing dataset.
3. Arabic web page analyzer: This tool extracts and evaluated the set of proposed Arabic Web spam features of Wahsheh et al. (2013b).

We analyzed the Arabic Web spam dataset using the set of proposed Web spam features which are presented in Table 1.

Our dataset in this study is evaluated against those listed Arabic content and link Web spam guidelines to define possible usages of spam techniques in Arabic Websites. In order to make the decision that a Website is a spam or not, we need to extract all features of the Web pages that composed that Website (not only the home page). For the spam Websites, some of their Web pages can use spam techniques, while the other Web pages are normal Web pages. So in order to identify a website as a spam Website we have to determine the percentage of spam Web pages within a given Website. In this study any Website is considered as a spam Website if the percentage of spam Web pages within the Web site is 70% or more.

Table 1 Arabic Web spam features (Wahsheh et al., 2013b).

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

For each one the 24 investigated Websites, we evaluated 100 Web pages. This means that we analyzed 2400 Web pages of 24 Arabic top Websites. It should be mentioned that we exclude all Arabic top Websites with trusted domains (i.e., .edu and .gov domains).

Table 2 shows a sample that is composed from twenty-four popular Web pages which is studied and evaluated in this study.

The common non Arabic spam Web pages are characterized by their long URLs, so spammers normally add many spam words to the spammed URLs (Gyongyi and Garcia-Molina, 2005). However, Table 2 presents a different case of common spam Websites, which shows that the popular Arabic Websites under test were characterized by their short URLs. These twenty-four Arabic spam Websites are identified by Alexa.com as popular Websites which appeared in the SERP by searching using popular Arabic words.

Table 3 presents another sample of popular Arabic Websites. These Websites are considered as suspected spam Websites, since they contain a high number of out-links and many images which are used to attract users to spammed Websites.

It should be noticed that not all Web pages that have a large number of images and outlinks are spam Web pages. However, this technique is used by a large portion of spammed Web pages. Therefore, the Web page that has a large number of images and out-links is considered a suspected spam Web page, and not identified for granted as a spam Web page. The content of these spam Web pages is usually different from the content of images they have. Therefore, the decision for these Web pages as a spam or not depends on the users' feedbacks.

Outlinks are links from a Web page to other Websites or Web pages. Therefore, spammers usually use outlinks to refer to other spammed Web pages. The outlinks are used to connect different Web pages to each other, but they are also used by Web search engines to compute the popularity of different Web pages. However, irrelevant links are usually considered as suspected spams.

Table 4 shows the number of Meta words in the spam Web page or its head particularly. Meta words are used to help Web

Table 2 A sample of popular Arabic Websites under test.

Web page with Short URL		
graaam.com	arabic.qiran.com	rjaah.com
damasgate.com	jiro7.com	iraq3.com
12allchat.com	sa-l.com	arabchat.net
iq29.com	kuwait29.com	newmar.net
ct-7ob.com	x333x.com	ksavip.com
hesn-3.com	bnatksa.com	drdsh.com
arabchat.com	safara.com	lo2l.net
qcat.net	newcoool.com	dardaasha.com

Table 3 Suspected spam Websites with their out links (external) and images.

Web page	Out-links	Web page	Images
Damasgate.com	142	hesn-3.com	74
hesn-3.com	143	jiro7.com	94
Arabchat.com	130	x333x.com	165
jiro7.com	96	Rajah.com	118
sa-l.com	328	iraq3.com	122
x333x.com	159	Newcoool.com	193

Table 4 Size of Spammed Meta element words.

Web page	Meta words	Web page	Meta words
Damasgate.com	51	Safara.com	33
12allchat.com	46	Rajah.com	101
hesn-3.com	91	iraq3.com	62
Arabchat.com	193	Arabchat.com	105
arabic.qiran.com	31	Ksavip.com	139
jiro7.com	117	lo2l.net	37
sa-l.com	43	Qcat.net	47
kuwait29.com	51	dardaasha.com	36

Table 5 Size of Suspected < title > element words.

Web page	Title words	Web page	Title words
12allchat.com	15	Kuwait29.com	8
Iq29.com	15	X333x.com	11
Ct-7ob.com	9	Iraq3.com	11
Hesn-3.com	8	Arabchat.com	15
Sa-l.com	11	Newmar.com	23
Drdsh.com	8	Ksavip.com	38

search engines to determine the nature of the Web page and its content. The role of using Meta words in different Web pages is exactly similar to the role of using keywords in research studies. Therefore, these Meta words should help to classify different Web pages. Web spammers may stuff their spam Web pages with many popular keywords, to make their Web pages relevant for most of the queries.

Fig. 3 shows the number of words within < title > element in spam Websites and non-spam Websites.

Increasing the number of words inside the < title > element will help the Web page to obtain a better PageRank score. Therefore it is known that the high number of words inside the < title > element may lead to the assumption that the Web page is a suspected spam Webpage, since spammers know and exhibit this type of behavior. This is known as the keyword stuffing technique which is used inside < title > to gain a high rank within SERP. The threshold to this is to be up to three times the original or the norm. If it exceeds three, there is a downturn in terms of visibility (Wahsheh et al., 2013b). Fig. 3 shows clearly that average Arabic/English word number inside < title > element in spammed Web pages exceeds its average counterpart within non-spam Arabic Web pages.

Table 5 shows the number of possible spam words in the titles of Web pages. While results showed that some Web pages have used spam techniques of all types, we can see that most of the popular or top ranked Web pages in Arabic use one technique or more.

Each one of the popular Arabic Websites that is used in this study can be classified either as an entertainment or social networking Web page. This may explain why administrators and Web masters of these Websites are not fully aware of ethics and used unethical techniques to improve the visibility of their Websites. Sometimes Web search engines consider the use of spamming techniques as unintentional or as unprofessional. Therefore, there is a need to enforce Webmasters and Web programmers to be Search Engine Optimization (SEO) certified.

The evaluation of a web page whether it is a spam web page or not is performed through our developed spam detection engine. This spam detection engine is filled with rules that will detect if any one of those spam behavior rules is applied to the web page and if so, it is classified as a spam page.

In this study we used the WEKA data mining tool, in order to summarize the evaluation of the spammed behavior of the top popular Arabic Websites (2400 Web pages) against the normal behavior of the normal popular Websites, which contains 2400 normal Web pages, that are available in the dataset of Wahsheh et al. (2013b).

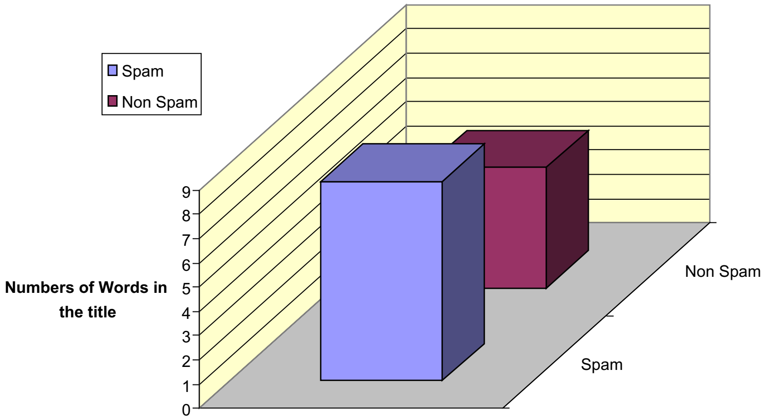


Figure 3 Content size of < title > element in spam and non-spam Websites.

Table 6 Accuracy information results using Naïve Bayes algorithm.

Class	True positive	False positive	Precision	Recall	*F*-measure	Receiver operating characteristic
Spam	0.918	0.47	0.662	0.918	0.769	0.908
Non- spam	0.53	0.082	0.867	0.53	0.658	0.908
Weighted AVG	0.724	0.276	0.764	0.724	0.724	0.908

Table 6 presents the summarization accuracy information results that distinguished spam and non-spam Websites, using Naïve Bayes algorithm. This algorithm is also used by Altwaijry and Algarny (2012) to detect different intrusions.

Table 6 shows that Naïve Bayes algorithm can distinguish the spam and non-spam Websites through the used Web spam features, which yields an accuracy of 71.875%.

5. Conclusion

Website masters and developers struggle to improve their Websites' popularity and visibility; such actions help to increase the value of the Websites and give them better values in terms of e-commerce, marketing, advertisements, etc.

In this paper, we selected most popular Arabic Web pages in the Middle East region according to Alexa.com ranking during 2012 fourth quarter. We evaluated those popular Websites against the possible usage of spam techniques. Results showed that the majority of those Web pages use spamming techniques with different levels and approaches. We noticed also that the majority of the popular Web pages in Arab region are either classified as entertainment or social media Web pages. We also focus on those Websites and exclude Websites of possible trusted domains such as: (.edu or .gov). However, this assumption, whether such trusted Websites, may have less usage of spam should be further investigated. Visibility to entertainment and social networks' Websites is very important. Spam techniques can be then used to increase such visibility.

The NB classifier is used to classify Web pages into Spam or non-spam. The performance metrics prediction, recall, *F*-measure, and the area under the ROC curve are measured to show the quality or accuracy of the predicted classification.

We believed however, that the classification of Web pages into Spam and non-spam is not yet mature, especially for Arabic Websites. There are some criteria that are not widely agreed upon to be considered as a spam behavior or not. In fact, search engines conduct some activities that are banned by themselves, if conducted by others, and hence classified as spam techniques.

References

Al-Kadhi, M.A., 2011. Assessment of the status of spam in the Kingdom of Saudi Arabia. J. King Saud Univ. Comput. Inf. Sci. 23, 45–58.

Altwaijry, H., Algarny, S., 2012. Bayesian based intrusion detection system. J. King Saud Univ. Comput. Inf. Sci. 24, 1–6.

Baeza-Yates, R., Ribeiro-Neto, B., 2010. Modern information retrieval: the concepts and technology behind search. Addison-Wesley Professional, Indianapolis, Indiana.

Berlt, K., Moura, E., Carvalho, A., Cristo, M., Ziviani, N., Couto, T., 2010. Modeling the Web as a hypergraph to compute page reputation. Inf. Syst. 35, 530–543.

Bhushan, B., Kumar, N., 2010. Searching the most authoritative & obscure sources from the Web. IJCSNS Int. J. Comput. Sci. Netw. Secur. 10, 149–153.

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

Chung, Y., Toyoda, M., Kitsuregawa, M. 2010. Identifying spam link generators for monitoring emerging Web spam. In: Proceedings of the 4th workshop on Information credibility WICOW '10, pp. 51–58.

Du, Y., Shi, Y., Zhao, X., 2007. Using spam farm to boost PageRank. In: The proceedings of the 3rd international workshop on Adversarial information retrieval on the Web AIRWeb '07. ACM, pp. 29–36.

Goodstein, M., Vassilevska, V., 2007. A two player game to combat Web spam. School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, pp. 1–22.

Gyongyi, Z., Garcia-Molina, H. 2005.Web spam taxonomy, In: Proceedings of the 1st international workshop on adversarial information retrieval on the Web, Chiba, Japan, pp. 1–9.

Jayanthi, S., Sasikala, S., 2011. DBLC_SPAMCLUST: spamdexing detection by clustering clique-attacks in web search engines. Int. J. Eng. Sci. Technol. (IJEST) 3, 4572–4580.

Kang, F., Liu, X., Liu, W. 2011.A personalized ranking approach via incorporating users' click link information into PageRank algoritm, In: International conference on energy systems and electrical power (ESEP 2011), Vol. 13, pp. 275–284.

Kerchove, C., Ninove, L., Dooren, P., 2008. Maximizing PageRank via external links. Linear Algebra and its Applications 429, 1254–1276.

Largillier, T., Peyronnet, S., 2011. Detecting Web spam beneficiaries using information collected by the random surfer. Int. J. Organizational Collective Intell. IJOCI 2, 1–17.

Li, D., Walejko, G., 2008. Splogs and abandoned blogs: the perils ofsampling bloggers and their blogs. Inf. Commun. Soc. 2, 279–296.

Moe, H., Walejko, G., 2011. Mapping the Norwegian blogosphere: methodological challenges in internationalizing internet research. Social Science Computer Review, 313–326.

Schwarz, J., Morris, M. 2011. Augmenting Web pages and search results to support credibility assessment, CHI 2011, Vancouver, BC, Canada, pp. 1–10.

Selvan, M., Sekar, A., Dharshini, A., 2012. Survey on web page ranking algorithms. Int. J. Comput. Appl. 41, 1–7.

Shen, G., Gao, B., Liu, T., Feng, G., Song, S., Li, H., 2006. Detecting link spam using temporal information. In: Proceedings of the sixth international conference on data mining pages ICDM '06. IEEE, pp. 1049–1053.

Wahsheh, H., Alsmadi, I., Al-Kabi, M. 2013a. Evaluation of Web spam behaviour on Arabic Websites popularity, In: Proceedings of the 6th International Conference on Information Technology, ICIT'13, Amman, Jordan, pp. 1–7.

Wahsheh, H.A., Al-Kabi, M.N., Alsmadi, I.M., 2013b. A link and content hybrid approach for Arabic web spam detection. Int. J. Intell. Syst. Appl. (IJISA) 5 (1), 30–43.

Zhou, B., Pei, J., 2009. Link spam target detection using page farms. ACM Trans. Knowl. Disc. Data (TKDD) 3, 1–38.