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

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.