Analyzing the Popular Words to Evaluate Spam in Arabic Web Pages

Heider A. Wahsheh  
CIS Department  
IT & CS Faculty  
Yarmouk University  
Irbid, Jordan  
heiderwahsheh@yahoo.com

Izzat M. Alsmadi  
CIS Department  
IT & CS Faculty  
Yarmouk University  
Irbid, Jordan  
ialsmadi@yu.edu.jo

Mohammed N. Al-Kabi  
CIS Department  
IT & CS Faculty  
Yarmouk University  
Irbid, Jordan  
mohammedk@yu.edu.jo

ABSTRACT
The extensive expansion and use of the Web and the Internet comes at the price of seducing a number of intruders to utilize the Web for destructive purposes. In the scope of Websites and Web pages, spammers try to inject their own content and pages in Web sites and search engine search results to be more visible to users and attract users to their Websites or products.

This paper analyses the behaviors of the spammers in the content-based Arabic Web pages, through analyzing the weights of the most ten popular Arabic words used by Arab users in their queries. The results show that the behavior of the spammers in the Arabic Web pages can be unique and distinguished in comparison to other languages. Decision Tree was used to evaluate this behavior and it obtains the degree of accuracy which is equal to 90%.

General Terms
Security, Legal Aspects.

Keywords
Arabic spam, content-based, spammer behavior, Arabic Keywords.

1. INTRODUCTION
The total of population in Arab countries is around 350 million people (5% of the World population), and the total of Arab Internet users is around 65 million users (3.3% of the total Internet users) [20].

The Arab world is witnessing at the present time an increase in Internet usage, which includes Internet browsing to retrieve, present, and traverse WWW information resources. The information and knowledge that stored in this giant reservoir are not organized, therefore most of Web users use search engines to retrieve the relevant portion of information and knowledge to their needs from this giant pool. Naturally the increase in Internet usage in the Arab world during the last five years leads to an increase in Web search. A conducted survey rank the Arab countries according to their penetration rate score in descending order as follows: United Arab Emirates, Bahrain, Qatar, Kuwait, Morocco, Tunisia, Saudi Arabia, Jordan, Lebanon, Egypt, Syria, Oman, Palestine, Algeria, Sudan, Libya, Comoros, Djibouti, Mauritania, Yemen, Iraq, and Somalia with the following Arabic speaking Internet users penetration percentages: 74.10%, 55.30%, 52.30%, 37.10%, 33.40%, 33.40%, 27.10%, 25.40%, 23.50%, 21.10%, 16.40%, 16.30%, 14.40%, 12.00%, 10.20%, 5.10%, 3.10%, 2.60%, 1.90%, 1.60%, and 1.00% [5]. This statistics indicates that the use of the Internet is flourishing rapidly in countries like United Arab Emirates, Bahrain, Qatar, Kuwait, while it’s creeping in countries like Iraq and Somalia which ranked the lowest [4].

The huge numbers of Web pages include various content of text, audio, video, and the combination of media features. This variety of Web pages types lead to attract more users to the Internet and search world to benefit from various interactive features such as: e-shopping, emailing, chatting and downloading media in different topics [20].

The percentage of Arabic content on the internet and within the free encyclopedia (Wikipedia) does not exceed 1% of the total Internet content in different languages. This percentage is less than the percentage of Arab Internet Users (3.3%), and less than the percentage of Arab population (5%). Arabic internet slang language constitutes around one third of the total Internet Arabic content, where the content of Arabic blogs and forums constitute 35% of this total [12, 20].

Most of the owners and Webmasters of commercial Websites attempt to make their Websites or certain Web pages ranked high in the different listings of Web pages presented by Web search engines as a response to different queries. These listings are also known as search engine results pages (SERPs), where any increase in the ranking of such Web site/Web page leads to increase owning company revenues, especially when a URL of Web page lie on the first SERP (top 10).

To achieve the goal of ranking high within SERP, owners and Webmasters of commercial Websites either tries to use ethical ways represented by what it is known as search engines optimization (SEO) techniques, or they try to adopt spamming techniques which considered unethical ways to rank different Web pages higher than they really deserve.

Web spam led to deteriorate the quality of search engine. There are three main types of Web spam techniques depending on their role: Content-based spam, Link-based spam, and Cloaking. Content-based spamming techniques aim to change search engine view of the content of Web pages adapting these techniques. These techniques includes, Keyword stuffing, Hidden/invisible text, Meta-tag stuffing, Article spinning, Scaper sites and Doorway pages [10, 15]. Link-
based spamming techniques aim to manipulate the hyperlinks structure, and point to irrelevant Web pages, to increase the rank of particular Web pages. Cloaking is the hybrid of content-based and link-based spamming; which is based on sending two different versions of the same Web page (URL), where one of these dedicated to search engine crawlers (spiders), and the other one is dedicated to different Web browsers [10].

Several efforts and studies have been made to detect content-based Web spam techniques, and few of these studies interested in Arabic content-based spam Web pages [3, 13, 22, and 23].

Only four previous studies are dedicated towards the detection of Arabic Web spam. Four of them dedicated for Arabic content-based spamming [3, 13, 22, and 23], where these studies are based on extracting certain features of each Web page in the dataset, to determine whether is a spam or non-spam Web page. This study is based on Keywords extracted by Google’s free Search-Based Keyword tool (SBK). SBK tool is released by Google originally to provide advertisers with a list of highly relevant keywords to be added to their sites in different natural languages, where these keywords are extracted from actual Google search queries using different natural languages [2]. The main concept of this study is based on the analysis of spammer’s behavior using the most popular Arabic keywords, which is different from the four previous studies related to the detection of Arabic content-based spamming [3, 13, 22, and 23]. The top 10 Arabic keywords extracted by SBK for all Arab countries are: ‘chat’ (دردشة), ‘Games’ (ألعاب), ‘YouTube’ (يوتيوب), ‘Facebook’ (فيسبوك), ‘University’ (جامعة), ‘Forums’ (منتديات), ‘Photos’ (صور), ‘Billiards’ (بلياردو), ‘Trab’ (طراب).

Table 1 shows the number of retrieved pages for the top 10 most popular Arabic keywords extracted by SBK tool in Arab countries.

<table>
<thead>
<tr>
<th>Arabic Query</th>
<th>SBK results</th>
<th>Translated Query</th>
<th>SBK results</th>
</tr>
</thead>
<tbody>
<tr>
<td>العام</td>
<td>4,090,000</td>
<td>Games</td>
<td>5,000,000</td>
</tr>
<tr>
<td>صور</td>
<td>2,240,000</td>
<td>Photos</td>
<td>201,000</td>
</tr>
<tr>
<td>_above</td>
<td>1,830,000</td>
<td>Songs</td>
<td>1,830,000</td>
</tr>
<tr>
<td>فيس بوك</td>
<td>1,830,000</td>
<td>Facebook</td>
<td>7,480,000</td>
</tr>
<tr>
<td>يوتيوب</td>
<td>1,220,000</td>
<td>YouTube</td>
<td>9,900</td>
</tr>
<tr>
<td>جامعة</td>
<td>1,220,000</td>
<td>University</td>
<td>1,220,000</td>
</tr>
<tr>
<td>دردشة</td>
<td>1,000,000</td>
<td>Chat</td>
<td>1,500,000</td>
</tr>
<tr>
<td>منتديات</td>
<td>550,000</td>
<td>Forums</td>
<td>550,000</td>
</tr>
<tr>
<td>طراب</td>
<td>165,000</td>
<td>Trab</td>
<td>12,100</td>
</tr>
<tr>
<td>بلياردو</td>
<td>201,000</td>
<td>Billiards</td>
<td>12,100</td>
</tr>
</tbody>
</table>

The conducted analysis by our team reveals that the computed weights of these selected Arabic Keywords within Arabic Web spam pages exceeds their computed weights in non-spam Web pages, where the Site Content Analyzer tool is used to compute keywords weights within all Web pages. Site Content Analyzer tool computes the weights of these keywords everywhere, regardless of the keyword position inside the HTML document.

The rest of this study is organized as follows: Section two presents a brief review of previous work related to content-based Web spam detection, section three describes our methodology to evaluate the spammers behaviors within Arabic Web pages using popular words. Section four presents the Decision Tree and the experiment results. Section five presents the conclusion and future work.

2. RELATED WORK

The Term Frequency–Inverse Document Frequency (TF-IDF) is a weight computed for each word in a collection of documents, which is usually used in information retrieval and text mining. This weight usually shows the importance for each word within each document in a corpus, and TF-IDF has many variants. High TF-IDF score implies a strong relationship between the word and the document the word appears in, and that means that if this word also appears in a query, the document in this case could be a relevant document for that query [14]. TF-IDF weighting scheme are often used by search engines as a central tool in ranking a document’s relevance to particular user query, so the spammers try to increase the TF-IDF scores for their Web pages, by including a large number of popular terms in a document, to gain a higher possible TF-IDF weights [10].

[17] Study presents the spammer behaviors by using the globally popular words, with meaningless role in the context of Web pages, to increase the rank for these spam pages with globally popular words. The C4.5 classifier achieved highest accuracy of 86.2% for detecting spam Web pages.

[6] Exploit the linguistic features of Web pages to find and distinguish between spam and non-spam Web pages. They applied several machine learning approaches to detect Web spam. The conducted tests results reveal the effectiveness impact of the linguistic features to identify spam Web pages.

The [19] study based on the linguistic features. They studied around 200 linguistic attributes, by using two tools: Corleone and general Inquirer natural language processing tools, for the Web pages in the two public Web spam corpora (WebSpam-UK2006 and WebSpam-UK2007). Preliminary analysis indicates that certain linguistic features can be used for Web spam detection.

While [1] study proposed a novel model called Witch to identify spam Web pages. The model utilizes both content-based features of Web pages, and the Web graph structure. The test results showed that Witch model improve the effectiveness to identify Spam Web pages.

[9] Show through the machine learning based classifier the capability to develop new anti-spam techniques. Therefore they proposed a strategy based on predicted spamicity of learning algorithms and hyperlink propagation to combat the Web spam. The preliminary tests reveal the effectiveness of this strategy.

The [7] research based on the content-based features and temporal features within Web pages, to enhance of the accuracy of spam classifier. They presented their new classifier using machine learning; which combine the current page content-based with temporal features, where the conducted tests on this proposed classifier using WEBSPAM-UK2007 dataset show 30% improvement.

The study [11] produce a novel way to benefit from Web navigation behavior to detect spam bots through machine learning algorithms.
This paper [11] study using a proposed set of user behavior to identify the spam bots. Web navigation behavior used to establish the set of features to be used in training dataset in Support Vector Machine (SVM) classifier. The results yielded an accuracy of 96.24%.

The [16] study proposed a system which can be adapted and evolved to different spamdexing techniques, since it is based on genetic programming. The proposed system has the capability to gain the best possible discriminating function. The tests yielded a 26% enhancement within the recall performance, 11% improvement within the F-measure performance, and 4% improvement within the accuracy performance relative to SVM.

In the [21] the researchers presented a novel content trust model for spam detection of Web pages, such as URL Terms in both title and content, surrounding Anchor text, the relation between title and the content, Meta Tags, internal and external links. The experiments applied on two datasets (English & Chinese) of Web pages in order to enhance their model. The tests results have proved the effectiveness of the proposed model.

The statistical analysis studied in [8], and used machine generated for spam Web pages. Their study applied on 150 million Web pages, using linkage structure, and content-based features. Their study develops techniques that can help to identify spam Web pages based on the replication of phrases.

Another novel model for spam detection was proposed by [18]. This model is not based on words like other models, but it is based on phrases, to enhance the similarity measures in the title and body of Web pages. Evaluation tests have yielded an accuracy of 94% and it has gains better results in terms of F-measure by 10% than existing anti-spam methods.

[22] produced the first Arabic Web spam detection approach, they have manually built a small Arabic Web Spam corpus, which contains around 400 spam Web pages [22]. Three classification algorithms were tested; Decision Tree, Naïve Base, and K-Nearest Neighbour (K-NN). In their preliminary study [22] found that the K-NN classifier is more accurate than the other two, with an accuracy of 96.875% for detecting spam Web pages.

In [13] the study followed the study of [22], and has presented new content-based features of the Web pages to improve the effectiveness of the Web spam classifiers; which has also used three classifiers (Decision Tree, Naïve Bayes, and LogitBoost). The results have yielded that the Decision Tree gets the best results.

In [3] the study overlaps with other two studies conducted by [22] and [13]. They proposed an evaluation model based on extended content-based features and a larger dataset relative to those used in [13, 22]. The test results of the proposed model achieved good results, and showed that the Decision Tree algorithm is the best to identify spammed pages with an accuracy of 99.3462%.

In [23] the authors proposed a novel Arabic content-based Web spam detection system, which based on the rule-based of the Decision Tree classifier of 15,000 Arabic spam Web pages. The conducted tests showed an accuracy of 83% in detecting Arabic content-based Web spam.

3. METHODOLOGY

This paper aims to evaluate the spammers’ behavior in the content-based spam Arabic Web pages, through analyzing the weights of the top 10 popular Arabic words using by both Arab users in their queries, and by spammers to increase the rank for their Web pages. This methodology is based on the following high level steps:

1. Building a spam corpus based on the top 10 most frequently used keywords by Arabic Web searchers.
2. Extracting a set of content-based features from the collected Arabic Web pages only for the top 10 most frequently used keywords by Arabic Web searchers, using Site Content Analyzer software1.
3. Use Weka data mining tool, to evaluate the spam techniques, using the Decision Tree classification algorithm.

4. SPAM COLLECTIONS AND CONTENT-BASED FEATURES

4.1 Data Collection of Web Spam

Our current paper presents a unique Arabic spam dataset. This peculiar Arabic spam dataset is different from those used within [3, 13, 22, and 21]. It includes around 3,500 Web pages.

The popular keywords used in this study are ‘chat’ (دردشة), ‘Games’ (ألعاب), ‘YouTube’ (يوتيوب), ‘Facebook’ (فيس بوك), ‘University’ (جامعة), ‘Forums’ (منتديات), ‘Songs’ (أغاني), ‘Photos’ ( صور), ‘Billiards’ (بلياردو), ‘Trab’ (طرب).

4.2 Relevant Content-Based Features

We used the Site Content Analyzer tool, to extract the content-based features from different parts of HTML pages, for the most popular words in the Web pages. The extracted features are:

1. The number of popular word and it is density inside <body>.
2. The number of popular word and it is density inside <title>.
3. The number of popular word and it is density inside the Heading (<h1>…<h6>).
4. The number of popular word and it is density inside Anchor text.
5. The number of popular word and it is density inside Alt text.
6. The number of popular word and it is density inside the Link.
7. The number of popular word and it is density in the Bold format.
8. The number of popular word and it is density inside Italic format.
9. The number of popular word and it is density inside Image Src.

1 www.cleverstat.com
10. The number of popular word and it is density inside Meta Description.

11. The number of popular word and it is density inside Meta Keywords.

12. The number of popular word and it is density inside the Comment.

13. The total number of popular word in all the previous part of HTML Web page, the total density, and the total weight.

14. The total number of Key phrases which contains the popular word, the total density, and the total weight.

5. EXPERIMENT RESULTS

Through our follow-up of the results generated through our methodology, we noticed that the spammer’s used the top frequently used Keywords by Arab Web searchers in their spam Web pages. According to Search Engine Optimization (SEO) techniques that the contents of elements like <title>, <h1>...<h6>, <strong>, <a>...etc always gain higher weights by search engines. Therefore Arab spammers in most cases attempt to use these top frequently used keywords within the above elements without adding any value to them, but the main goal of this step was to deceive search engines and let a Web page gain a higher rank within SERP than it really deserves. The use of Keywords stuffing was evident in Arabic spam Web pages, where most frequently used Arabic keywords are used beside their meaningless same size Latin counterparts, where each in these two words (Arabic and Latin) share the same key on the bilingual Arabic/English keyboard.

Figure 1 shows an example of spammer’s behavior with ‘chat’ (تشات) word, used by Site Content Analyzer.

<table>
<thead>
<tr>
<th>Count</th>
<th>Weight</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>ahj</td>
<td>24</td>
<td>218.00</td>
</tr>
<tr>
<td>الأعمال</td>
<td>10</td>
<td>242.00</td>
</tr>
<tr>
<td>chat</td>
<td>9</td>
<td>222.00</td>
</tr>
<tr>
<td>hgag</td>
<td>6</td>
<td>24.00</td>
</tr>
</tbody>
</table>

Figure 1. An example of spammers’ behavior.

Figure 1 shows the actual weight for ‘chat’ (تشات) word is equal to 24 + 9 = 33, which is the summation of two words ahj and ‘chat’ (تشات). These two words ahj and ‘chat’ (تشات) are not semantically related, not related together in the meaning, they only share the same keys in the Arabic/English keyboard.

Figure 2 presents an example of normal case of Key phrases, with non-spam Web pages.

<table>
<thead>
<tr>
<th>Weight</th>
<th>Court</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>شات الإدرار</td>
<td>188.6</td>
<td>48</td>
</tr>
<tr>
<td>شات الإدرار المواقع</td>
<td>164.37</td>
<td>9</td>
</tr>
<tr>
<td>المواقع شات الإدرار</td>
<td>146.84</td>
<td>9</td>
</tr>
<tr>
<td>شات الإدرار chat</td>
<td>140.37</td>
<td>29</td>
</tr>
</tbody>
</table>

Figure 2. An example of key phrases in non-spam Web pages.

In example 2 we have two words, one in Arabic (شات) and it is translation in English (Chat).

The conducted tests showed the accuracy of 90% for detecting spam Web pages.

5. EXPERIMENT RESULTS

EXPERIMENT RESULTS

5. EXPERIMENT RESULTS

5. EXPERIMENT RESULTS

Figure 2. An example of key phrases in non-spam Web pages.

Figure 3. An example of spam behavior in the Key phrases with ‘chat’ (تشات), word.

In example 2 we have two words, one in Arabic (شات) and it is translation in English (Chat), which have a meaningful content in Web page. Each word has an independent weight, and this indicates to the normal state of Web pages.

In example 2 we have two words, one in Arabic (شات) and it is translation in English (Chat), which have a meaningful content in Web page. Each word has an independent weight, and this indicates to the normal state of Web pages.

Figure 3 shows how spammers use Arabic Key phrases in their Web pages. Spammers add their meaningless Latin words such as the Latin word (ahj) and the Arabic popular word (شات ‘chat’), where each corresponding two letters shares the same key on the bilingual Arabic/English keyboard. The weight of key phrase in this case will be duplicated.

Figure 3. An example of spam’s behavior in key phrases.

We used the Decision Tree (J48 in Weka) to evaluate the spam’s behavior in the collected Arabic Web pages, where the Decision Tree yielded 90% of accuracy.

6. CONCLUSION AND FUTURE WORK

The spammer’s tries continually to manipulate the content and link structure of Web pages to gain higher rank than the actual Web pages deserve. Three studies interested in Arabic content-based Web spam extracted a number of content-based features, and applied different algorithms to detect content-based Web spam.

In this paper, we analyzed the spammers’ behavior with the top 10 Arabic keywords used by Arab users in their queries, using SBK tool for all Arab countries. We collected a new Arabic spam corpus; which is mainly concerned with the content-based features for the top 10 Arabic keywords, and computed the weights of these words in everywhere inside the HTML document. Decision Tree classifier was used to evaluate the behavior of the spammers with the Arabic popular keywords, and the results showed that the spammers used unique Keywords stuffing techniques dedicated to Arabic language. The conducted tests showed the accuracy of 90% for detecting spam Web pages.
We plan to increase our spam dataset, and increase the content-based features to get more accurate results, using different algorithms to obtain the best Web spam detections algorithms. And also we plan to study the other types of Web spam such as link spam and cloaking to build a Web spam detection system for these types of spamming techniques.

7. REFERENCES