International Journal of Hybrid Information Technology
Vol.6, No.6 (2013), pp.109-126
http://dx.doi.org/10.14257/ijhit.2013.6.6.10

Detecting Arabic Cloaking Web Pages Using Hybrid Techniques

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

[1]*Computer Science Department, College of Computer Science*
King Khalid University, Abha, Saudi Arabia
[2]*Faculty of Sciences and IT, Zarqa University, Zarqa, Jordan*
[3]*Information Systems Department, College of Computer & Information Sciences*
Prince Sultan University, Riyadh, Saudi Arabia

heiderwahsheh@yahoo.com, mohammedk@zpu.edu.jo, ialsmadi@cis.psu.edu.sa

Abstract

Many challenges are emerging in the every day expanding Internet environment, whether for the Internet users or the Web sites owners. The Internet users need to retrieve the high quality relevant information which are relevant to their queries within a short period of time, in order to be a regular users who satisfied by search engine performance. While the Web site owners aim in most cases to increase the rank of their Web pages within SERP to attract more customers to their Web sites, and consequently gaining more visits, which in turn means more revenues.

The top rank of the Web pages within SERPs, is very important to the e-commerce and commercial Web pages. The owners of Web sites can attract more visitors to their Web pages, and gain more revenue, through Pay Per Click when their pages appear in the top results of SERPs. This paper proposed new approach of Arabic Web spam detection, dedicated with the cloaking Web pages, using hybrid techniques of content and link analysis. The proposed detection system built the first Arabic cloaking dataset contains around 5,000 Arabic cloaked Web pages. The proposed system extracts all possible rules from HTML element to monitor the cloaking behaviors, and then used three classification algorithms (K-NN, Decision Tree, and Logistic Recognition) in the experimental tests. This novel system yielded a high accuracy results with an accuracy of 94.1606% in detecting cloaking behaviors in Arabic Web pages.

Keywords: *Arabic Web spam, content-based, link-based, Arabic cloaking Web spam*

1. Introduction

Web is dynamic, since each day a huge number of Web documents are added to the Internet. Web is semi-static, since the volume of daily Web documents added to the Internet represents a relatively small fraction of the total volume of the Internet. Web Search engines represent a major outlet to access the Internet world, information, or documents, through the world. The main two characteristics distinguish features of good search engines are the speed and effectiveness, beside comprehensiveness which means the coverage of new added materials to the Internet. The main goal of these search engines is to display the largest percentage of URLs of relevant Web documents inside Search Engine Results Page (SERP).

Web spam or spamdexing is an illegal technique is (which a portmanteau of spamming and index), which aims to increase the rank of Web pages and Web

ISSN: 1738-9968 IJHIT

documents by deceiving Web crawlers. The description of Web crawler is usually based on the manipulation of content and link features of the spammed Web documents [1, 2].

Web spam uses many methods that manipulate the link and content features of the Web pages, such methods use:

- Invisible text: Web authors used the same color for the text and the background to hide the text manipulation of background colors in Web pages [3, 4, and 5].
- Keyword stuffing: This method is based on duplicating some Keywords or phrases, or inserting large number of unrelated words in the weighting tags of Web pages [3, 4].
- Tiny text: This method is based on inserting the keywords and phrases in a very small font and spread all over the Web pages, these tiny texts are not seen by the Internet users [3].
- Internal links and External links: Illegal exchanges of the links cause a manipulation in the Web pages ranks such as: Article spinning, Scraper sites and Doorway pages [3, 6].

The goal of these techniques is to make the spam Web pages as normal Web pages, and to attract more Internet users to visit spam Web pages [7].

Some of the owners of Arabic Web sites use the spam techniques and methods, which violate Search Engine Optimization (SEO), Search Engine Marketing (SEM), and Banners advertisements in order to rate their Web pages higher than they deserve. Usually spammed Web pages are characterized by their low information quality, and the crowdedness of advertising content and links, which deceive the search engine and thus lead to irrelevant information that does not match users' queries [1, 8].

The Arabic Web spam is considered as a part of the general Web spam. Some common features such as: Number of words in the Web pages, number of words in the title, number of popular words in Web pages, number of internal, external and redirected links is traced in both the Arabic and the general Web spam [9]. Yet, the spammers in Arabic Web pages deal with some special considerations that are related in particular to the Arabic language model.

Although we conducted many studies to discover Web spam with different techniques, finding an optimal effective solution for Arabic Web spam problem is still represent a big challenge to researchers in this field. Due the few number of researches interested in detecting Arabic Web spam, and the lack of reference collections in this field. This paper consider as a first study that focus in detecting the cloaking behaviors in Arabic Web pages. The key is to understand the methods and techniques which are used by spammers in Arabic cloaked Web pages. The next chapters of this paper are organized as follows: Chapter two presents' related studies to Web spam detection, chapter three presents the main weighting method and ranking algorithms for the Web pages, chapter four shows the proposed methodology, chapter five presents implementation and experimental results, chapter six described the evaluation of our proposed system. Last but not least Chapter seven presents the conclusions and future work.

2. Related Work

There are three main types of Web spam: content Web spam, link Web spam, and cloaking Web spam. Web spammers use these three types together in a single Web page or they can use any type they like.

This section divided into three subsections, each one focused in one of the main spam types as following:

2.1 Content-based Web spam Detection

Content Web spam refers to any manipulation process that aims at changing the content of Web pages, by using some spamming techniques, such as: Keyword stuffing inside any of the following tags: the <body>, <title>, Headers <h1>...<h6>, , and <Meta> tags, beside using tiny and invisible text [4].

In their study [10] proposed unsupervised method to detect spam documents from a given collection of documents, by using the string equivalence relations. The unsupervised method presented as scalable and language independent on many Web documents in Japanese language.

[11] Explored measuring framework for poor quality search results caused by the Web spam problem. About 80 million Web pages from UK2007 WEBSPAM were indexed by one machine. The evaluation method presented a sensitive difference between baseline and filtered rankings.

In [12] the authors built the first Arabic corpus of spammed Web pages contains around 400 Arabic content-based spam Web pages. A set of content-based features extracted as a first features related with Arabic Web spam. We used three classifiers; Decision Tree, Naïve Bayes, and K-Nearest Neighbour (K-NN). The results yielded that the K-NN is the best classifier in detecting Arabic content-based Web spam when K=1. While [13] proposed new groups of content-based features, and increase the size of the Arabic Web spam corpus. Three classifiers were tested (Decision Tree, Naïve Bayes, and LogitBoost). The conducted results showed that the Decision Tree is the best classifier in detecting Arabic content-based Web spam. In [14] the authors merged the two previous works, increase the size of spam corpus, where the number of Arabic Web spam is equal to the number of Arabic non Web spam. Proposed new content-based features, which mainly focused in key stuffing techniques. Another three classifiers used (Decision Tree, LogitBoost, and SVM). The results of the conducted tests presented that the Decision Tree is the best classifier to identify Arabic Web spam with an accuracy of 99.3462%.

In their study [15] built large Arabic content-based spam corpus containing 15,000 Arabic spammed Web pages, collected through customized Web crawler. Manual classification were used to label the Web pages in the spam corpus. Three of spam percentages (1%, 15%, and 50%) were used within three classifiers (Naïve Bayes, Decision Tree, SVM, and K-NN). The results showed that the Decision Tree is the best classifier with an accuracy of 99.96% in detecting Arabic content-based Web spam.

In [9] three of machine learning algorithms were used to detect the content-based Arabic Web spam. [9] built two corpuses of spammed Web pages; the first is extended Arabic Web spam, and the second dedicated with non Arabic Web spam. Using (Naïve Bayes, and Decision Tree) classifiers, the proposed content-based features showed that it suitable for detecting Arabic and non Arabic spammed Web pages.

[16] Studied the behaviors of spammers within the top ten Arabic words that used in Arabic queries using SBK tool for all Arab countries. A new spam corpus built which

concerned with the top ten Arabic keywords. The conducted tests monitor the spammer's key stuffing techniques with these popular keywords. The results presented the accuracy of detecting content-based Arabic Web spam with 90%, using Decision Tree classifier.

[17] The authors built Arabic Web spam detection system, using the rules of Decision Tree classification algorithm. They depend on the dataset of 15,000 Arabic spam Web pages. The proposed system yields an accuracy of 83%.

2.2 Link-based Web spam detection

Link spam which is based on the manipulation of the link structure of Web pages, such as: inserting irrelevant links to point to other Web pages in illegal techniques. There are two main techniques used by link-spam: Link hijacking is the most popular technique and uses for link-spam, which is based on using many links to point a target spam Web page called controlled page to arise its rank [18]. The second technique is the Honeypots: This is an indirect spam way, either entrapping a reputable Web page by inserting spam links, or using many links pointing to each other as controlled pages [5, 18].

The study of [3] focused on detecting link-based Web spam, and ignored the content-based Web spam features. They calculate the scores of the set of link-based features for each Web page, and applying the rank propagation and probabilistic over the Web graph structure. They built the classifier which tested on the large Web link spam dataset. The tests used the ten-fold cross-validation, and the best classifier detect 80% of spam Web sites with only 2% as a false positives rate.

The study of [18] analyzed the page quality, and extract the link credibility through three distinct features. Semi-automatically technique is used to: Evaluate different pages link credibility, allow the personal user to assess the link credibility, and propose CredibleRank. The proposed CredibleRank algorithm is based on credibility metrics and quality of page scores superior to PageRank and TrustRank algorithms.

While the study of [19] used an online learning algorithm to monitor the host which can generate the link-based Web spam. Those researchers investigate on the Web spam seeds and extracted the set of link-based features which affect the PageRank score. The experiment tests used the archives of Japanese Web pages, and yields the precision between 56% and 73% and yields F-measure between 0.54 and 0.68.

The study of [20] conducted a series of studies related to Web spam in Web 2.0 platforms, started with [21] in which the spammers insert their spam URLs in popular Web sites such as: social networking service, and Yahoo news. The spammers used spam methods such as: Web 2.0 spam which duplicates the amount of message spam, and Web spam bots which increase the spread of the spam content. While [21] proposed the Honey spam 2.0 tool which monitors the Web bot behavior. In their study [22] continued the interesting on the spam bots, they used the action time and action frequency to detect the spam bots. The results showed that the accuracy enhanced and reaches 94.7%.

The study of [23] explored an automated supervised machine learning technique to detect the spam bots inside 2.0 platforms. The new approach focuses on the navigation behavior, and compares between users and spam bots behavior. The conducted tests used Matthew Correlation Coefficient method, and have yields an accuracy of 96.24%.

The study of [24] focused on the Arabic link-based spamming techniques. They built link-based spammed corpus includes 3,000 Arabic spammed Web pages. They explored the irrelevant exchanging of external and internal links in the Arabic spam farm.

Decision Tree, and Naïve Bayes classifiers used to evaluate the Arabic link-based Web pages, the results showed accuracy of 91.4706% with Decision Tree as the best classifier in detecting link-based Arabic Web spam.

2.3 Non Arabic cloaking and hybrid Web spam detection

The last type of Web spam is cloaking. This type is based on the simple idea of producing two different versions for each spammed Web page; the difference comprises the content and quality. One of the two versions is meant to be with high and valuable information quality, and is sent to the Web crawler to achieve a high rank. While the second version is meant to be a spam Web page, and is sent to the user browser [25]. Sometimes cloaking Web spam is viewed as a hybrid type of the two previous types (content and link spam), due to the similarity in the features with the content Web spam from one side, and the similarity with redirection which used in cloaking with the linked Web spam from the other side.

The study of [26] reported cloaking and redirection techniques as important spamdexing techniques. It produces a realistic view of content-based and link-based methods to detect cloaking and link redirections, through the computation of three different copies for each Web page. The analysis results estimated (through two used datasets) that 3% of the first dataset and 9% of the second dataset used the cloaking technique to increase the rank of their Web pages.

In their study [27] showed the ratio of cloaking SERP which based on the query properties such as: link popularity and monetizability. They proposed new metrics for detecting cloaked Web pages through normalizing the TF between multiple downloaded Web pages versions. The experiments claimed using 10,000 search queries and 3 million related SERP. The results presented that 73.1% of the cloaked popular search Web pages are spam, and around 98% of the cloaked monetizable Web pages are spam.

In their study [28] presented a combination of link-based, and content-based spam detection features. The spam detection system divided into three phases; initially clusters the host graphs and labels all the hosts, then predicts labels to neighboring hosts, and finally uses the predicted labels as new features and retests the classifier. The researchers found that the linked hosts belong to the same cluster: either both are spam or both are non spam. The results of the best classifier showed an accuracy of 88.4% with 6.3% false positives.

In the study of [29] the authors claimed that they have performed the larger characterization content-based and HTTP analysis methods on the 350,000 Web documents. The analysis of content showed the duplication of the information content and URLs redirections. The analysis of content-based spam Web pages divided the Web documents into five main categories: Advertising Farms, Parked Domains, Advertisements, Pornography, and Redirection. While the link-based analysis performed, showed that the spammers used narrow IP addresses ranges for the spam hosting.

The study of [30] proposed new technique based on a new machine learning approach, which can detect both link-based features, and word-based features, which

extracted using the combinatorial feature-fusion method. They used many human-engineered features constructed from the raw data, and used semi-supervised learning to classify the unlabelled test Web pages. The results showed the high effectiveness of semi-supervised learning, and the combinatorial feature-fusion method.

The URL redirection is considered as a one of the cloaking styles. In their study [31] reported the employment of the redirection on the spam links. The study found that the redirection is widely used, about 40% of all links for different goals. Several JavaScript of URL, Meta and server sides' redirections were detected in the spam Web pages as internal and external links. The experiment results applied on the legitimate Alexa, UGA, and blogs datasets, showed that the percentage of using external redirection techniques by spammers is 46.81%, and the percentage of using external redirection techniques by legitimate sites is 53.19 %.

The study of [27] presented the cloaking Web spam in form of spamdexing redirections with false content to Web crawler for indexing purposes, while the irrelevant content sent to user browser. They studied famous JavaScript redirection techniques, which are stronger than the static analysis and static feature based detection systems. Their research found that the use of light weight JavaScript parsers is effective to predict the redirection of spam behaviors.

A study by [32] presented a novel algorithm; called Web spam Identification Through Content and Hyperlinks (WITCH) which aims to learn the Web spam detection techniques on both Web sites and Web pages level. The Witch algorithm takes the advantages of the content-based features and the Web graph structure. The authors claimed that Witch algorithm outperformed the other algorithms in the scalable, efficiency, and the state-of-the-art accuracy using SVM which achieved accurate results in detecting Web spam.

The study of [33] proposed language model approach which extracted a combination of content-based and link-based features from two popular spam datasets (Webspam-Uk2006 and Webspam-Uk2007). Kullback-Leibler (KL) divergence was applied on the spam Web pages to characterize the relation between the two linked Web pages. The proposed model has improved the F-measure of Webspam-Uk2006, and Webspam-Uk2007 to about 6% and 2% respectively.

In their study [34] the authors explored the survey of different spam methods, and filtering algorithms. The existing solutions were dedicated to content-based, link-based, and non-traditional data (*i.e.*, user behaviors, clicks, and HTTP sessions) Web spam detection. Their anti-spam algorithms provided high successful Web spam detection results with an accuracy of 90%.

3. Proposed Frame work

The frame work of Arabic cloaking Web spam detection system is summarizing in the Figure 1.

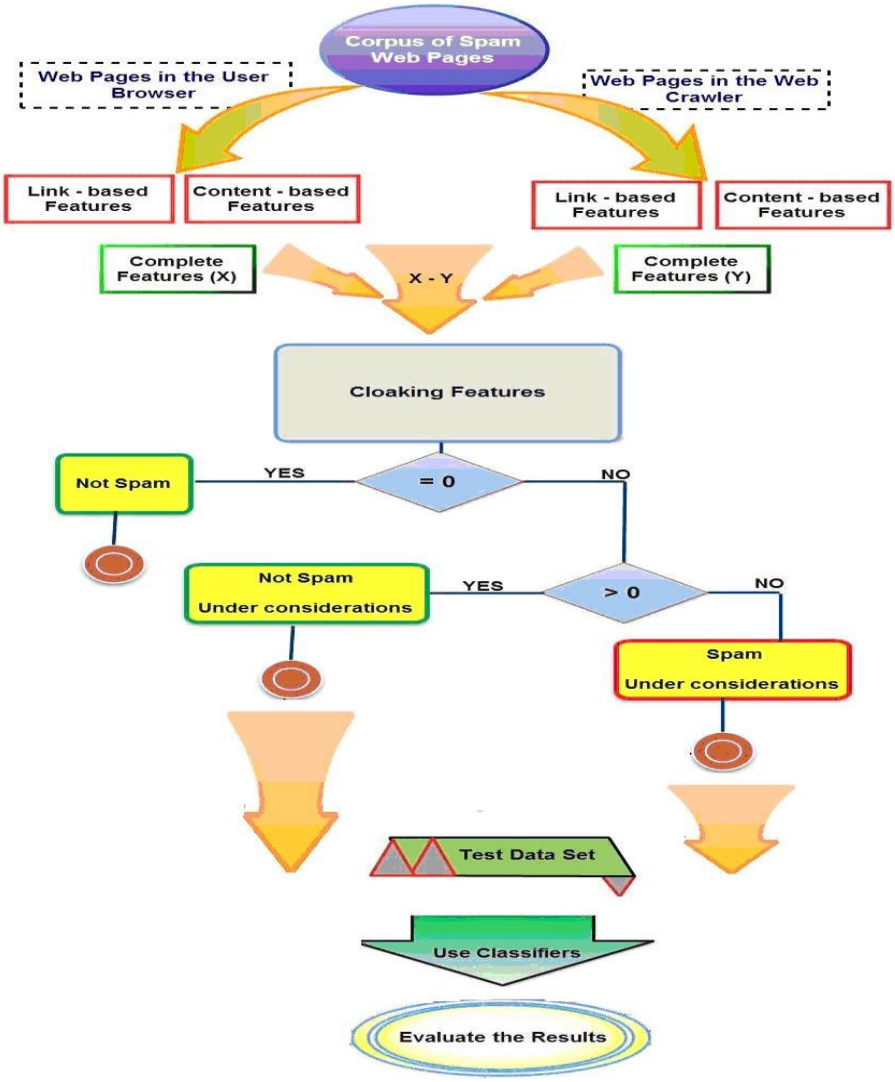

Figure 1. Main steps to detect Arabic cloaking Web spam

The Frame work of detecting cloaking Arabic Web spam divided into the following seven main steps:

1. Use Web crawler; which collect the Web pages, to build the cloaking Web spam dataset containing 5,000 Arabic Web pages. 1,000 of them to evaluate the proposed Arabic cloaking Web spam detection system and parsed all the hyperlinks, and the content in it.
2. Develop a Web page analyzer which extract the cloaking features, through analyze the content/link features, and then find the difference between the two copies (user browser and Web crawler) of each Web page in every HTML elements in the Web pages.
3. Identify the cloaked behaviors in each Web page with different consideration.

4. Evaluate the frame work of detecting cloaking Arabic Web spam, using the three classification algorithms Decision Tree, Logistic Regression, and *K-NN*.

4. Cloaking Features Extraction

Cloaking spam Web pages is based on the basic idea of producing two different versions of each Web page. The difference between them affected by the factors of content-based, link-based features, and the quality of the Web page. The high quality version appearing on the Web crawler to get the rank that actually contrary to the quality version appearing in the user browser [25].

In this paper we used the improved Web crawler to collect the first Arabic cloaking Web spam dataset, which contains around 5,000 Arabic Web pages with cloaked behavior. The cloaking Web page analyzer find the difference between the content-based and link-based features which appears in the user browser; and the content-based and link-based features which appears in the Web crawler.

Figure 2 presents the content and link features which extracted from user browser and Web crawler to find the cloaking features:

Cloaking Feature	Description
1. The number of meaningless e English/Arabic words in the <body> elements of Web pages.	Spammers used this technique widely; meaningless English/Arabic words provide an important weights that increase the rank of the spammed Web pages [16].
2. The total number of characters (Arabic, English, or Symbol) in all <Meta> elements of a Web page under consideration.	If we have *n* as a number of <Meta> elements in specific Web page, the Web page analyzer compute the number of characters in each <Meta> as a independent element of other <Meta> elements in this specific Web page.
3. The total number of Arabic/English words in the all <Meta> elements in a specific Web page.	If we have *n* as a number of <Meta> elements in specific Web page, the Web page analyzer compute the number of Arabic/English words in each <Meta> as a independent element of other <Meta> elements in this specific Web page.
4. The total number of Symbol words or characters in the all <Meta> elements in a specific Web page.	The symbol words or characters are composed of letters, unique characters, and punctuation marks that may appear in some Arabic Web pages. Therefore it is considered as a candidate to be one of Arabic spammed features. So we extract the total number of characters of the symbols in all <Meta> elements of the Web page. Our Arabic content/link Web spam detection system will check if the spammers use those strange symbols to increase the rank of their Web pages, or not.
5. The total number of Arabic, English, or Symbol words or characters in <body> element or in a specific Web page.	Content-based feature.
6. The minimum length of Arabic, English, or Symbol word inside the <body> element, or in a specific Web page.	We assume that the minimum word consist of three characters.

7. The maximum length of Arabic, English, or Symbol word inside the <body> element, or in a specific Web page.	Content-based feature.
8. The Average length of Arabic, English, or Symbol word inside the <body> element, and in a specific Web page.	Content-based feature.
9. The total number of <Meta> element inside a specific Web page.	Spammers used many <Meta> elements, in order to use the key word stuffing in these elements.
10. The Web page size in Kilo bytes.	Spammers try to decrease the size of spammed Web pages to attract users to access in these Web pages.
11. The total number of characters with the URL.	Spammers try to insert many spam words in the spam URLs, to attract the users to use these URLs.
12. The complexity factor of Web page within lexical density inside the <body> element, or in a specific Web page.	Content-based feature.
13. The total number of Arabic/English/Symbol words inside the <title> element.	Content-based feature.
14. The size (Kilo bytes) of the hidden text inside the <body> element or in a specific Web page.	The spammers try to trick the search engines to see links and the content that are not visible to normal users. This can be done through embedding them in very small pictures, or using tiny text font, or using the same color as the page background.
15. The total size (Kilo bytes) of compressed files inside the <body> element or in a specific Web page.	Content-based feature.
16. The total number of Arabic/English/Symbol words without repetition inside the <body> element or in a specific Web page.	Content-based feature.
17. The total number of image/images inside a specific Web page.	Spammers try to attract the users through using large number of meaningless images in their spam Web pages. They use these images to increase the traffic of visitors, so they can get more revenues.

18. The total number of links image/images inside a specific Web page.	Content-based feature.
19. The total number of the most popular Arabic words inside a specific Web page.	We based on the most popular Arabic words, and the English words that mentioned in [16].
20. The number of external links in specific Web pages.	Link-based feature.
21. The number of internal links in specific Web pages.	Link-based feature.
22. The total number of links (the internal and external) in specific Web pages.	Link-based feature.
23. The total number of characters in the URL	This feature can be considered as the same URL feature in the content, but it can be different when we have a redirected Web page.
24. The total number of broken links.	Link-based feature (decrease the PageRank).
25. The total number of redirected links in specific Web pages.	Link-based feature.
26. The total number of links without anchor text in specific Web pages.	Link-based feature.
27. The total number of anchor text without links in specific Web pages.	Link-based feature.

Figure 2. Web spam content and link features

Figures 3 presents' two versions of the same Web page as an example of cloaking Arabic spammed Web page.

Figure 3. Example of two version of Arabic cloaking Web page version

Figure 3 (left) shows the content that appears in the user browser version, while Figure 3 (right) presents the content appearing in the Web crawler for the same Web page.

5. Experimental Results

5.1 Cloaking Detection Classifiers

Three classifiers (Decision Tree, Logistic Regresion, *K Nearest neighbor (K-NN)*) given in Weka applied on the Arabic cloaking spam dataset. We have several evaluation metrics which validates the three classification algorithms, such as:

1. Accuracy: Represents a fraction of training documents that assigned to the correct class by the classifier [2].
2. Error: Represents a fraction of training documents that assigned to the incorrect class by the classifier [2].
3. Precision (*P*): Represents a fraction of dividing the number of relevant retrieved documents over the total number of retrieved documents [35].
4. Recall (*R*): Represents a fraction of dividing the number of relevant retrieved documents over the total number of relevant documents [35].
5. True Positive (*TP*): Represents the number of items correctly labeled as belonging to the positive class [35].
6. False Positive (*FP*): Represents the number of items correctly labeled as belonging to the negative class [35].
7. *F-measure*: Is an accuracy measure that combine both the precision and recall values. The traditional *F-measure* formula is:

$$F - Measure = \frac{2PR}{P + R} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where *P* is the Precision; *R* is the recall [2].

8. Receiver Operating Characteristic (ROC), or ROC curve: This curve depicts the performance of a binary classifier. It is plotting the fraction of True Positives Rate vs. the fraction of the False Positives Rate.

In the Arabic cloaking dataset we used three different spam percentages (2%, 30%, and 40%), which present the percentages group in the accuracy values of Arabic cloaking detection.

The Logistic Regression is applied on the three spam percentage groups to classify cloaking dataset. Applying it on the three spam percentage groups yield accuracies of 97.9933%, 87.3737%, and 86.1644% respectively. Table 1 shows the results of the accuracies and errors of using Logistic Regression within cloaking spammed Web pages.

Table 1. Cloaking logistic regression results

Spam Percentage Group	Accuracy	Error
2% spam Group.	97.9933%	2.0067%
30% spam Group.	87.3737%	12.6263%
40% spam Group.	86.1644%	13.8356%

The results shown in Table 1 indicate that we have not good percentages to detect the spam cloaking, on the three different percentages of spam. Table 2 shows the accuracy and error results of applying *K-NN* on the cloaking dataset. Applying the *K-NN* on the three spam groups, yield accuracy results of 98.3437%, 96.1014%, and 89.4434% respectively.

Table 2. Cloaking *K-NN* (IBK) results (*K*=1)

Spam Percentage Group	Accuracy	Error
2% spam Group.	98.3437%	1.6563%
30% spam Group.	96.1014%	3.8986%
40% spam Group.	89.4434%	10.5566%

The results shown in Table 3 indicate that the three percentage groups of spam dataset gain a high percentage to detect Arabic cloaking spammed Web pages.

Finally we applied the Decision Tree classifier on the three different percentages of spam, and it yields accuracies of 99.8174%, 96.1014%, and 89.4434% respectively. Table 3 shows the results of accuracies and errors.

Table 3. Cloaking decision tree results

Spam Percentage Group	Accuracy	Error
2% spam Group.	99.8174%	0.1826%
30% spam Group.	96.1014%	3.8986%
40% spam Group.	89.4434%	10.5566%

Table 3 shows the superiority of the Decision Tree when used the spam percentage of 2%.

In our spam dataset, we were considering the Web pages as a cloaking type if the Web page has a cloaking behavior in any of its HTML elements. So the extracted rules must check of every HTML element if it uses cloaking behavior or not.

5.2 Cloaking Results Analysis

After we extracted the cloaking features using cloaking Web page analyzer, we found that if difference result of the cloaking features is equal to zero. This means that we have not any differences between the user browser version and the Web crawler version. So the Web page under consideration is a non spam Web page.

If we found that the cloaking features have negative numerical results, this means that the Web crawler version has more content and links than the user browser version. So this indicates that we have a spam behavior, with some exceptions for external links and broken links. We need first to identify the thresholds that determine if the difference between the two versions (the user browser version and the Web crawler version) is significant to identify a spam behavior or not. The thresholds depend on the contents/links of HTML elements, so if we have a duplicate in the content and high number of additional links, this means that we reach the threshold. While if the difference in the content/link features between the user browser version and the Web crawler version does not has any duplication, or additional links. This means that we have not reached the thresholds of a spam behavior.

High number of broken links and external links in the Web crawler version means that the Web page will not get a high rank. This contradicts to the spam behavior which tries to increase the rank of Web pages. So if we found that we have negative numerical results with the broken links and external links in the cloaking features, this means that we have non spam behavior. This may considered as an errors in designing the Web pages by Web masters.

If the cloaking features have positive numerical results; this means that the user browser version has more content/links than the Web crawler version. So we have a non spam behavior, with some exceptions in external link and broken links. Increasing number of broken links and external links in the user browser version means that the version of the Web page that sent to the Web crawler does not contain the broken and external links. So these Web pages get a higher rank than it really deserve. Thus, the users will suffer from these spammed Web pages.

5.3 Cloaking Detection System

The results of the classified cloaked Web pages recommended that it is necessary to check every cloaking feature in the Web page, when we want to detect the cloaking behavior.

To extract all the cloaking features we need to extract all content/link features then find the difference between the two copies (user browser and Web crawler) of the cloaked Web page. Then the numerical values of cloaking features determine if we have a spam behavior or not.

Figure 4 shows the Algorithm of Arabic cloaking Web spam detection system.

Algorithm **Arabic Cloaking Web spam detection system.**

Input: **List of URLs stored on a text file (CloakingWebspam.txt).**

Output: **Table of the number of Web page (content-based and link-based) features, stored in the database of the Arabic Web spam detection system.**

```
BEGIN
  WHILE NOT EOF (CloakingURL.txt)
        Read the URL of a Web page.
        Download a Web page.
        Compute link/content features in the user browser.
        Compute link/content features in the Web crawler.
        Compute the difference values of link/content features (user browser-Web crawler).
        Make a decision of non spam/ or true percentage of spam.

  END WHILE
END
```

Figure 4. Arabic cloaking Web spam detection system.

6. Evaluate the Proposed Arabic Cloaking Detection System

We used the test dataset which consists of 1,000 Arabic spammed Web pages to evaluate the Arabic cloaking Web spam Detection System; the results yields 94.1606% accuracy in detecting cloaking Web spam.

Figure 5 shows the evaluation process of our Arabic cloaking Web spam detection system.

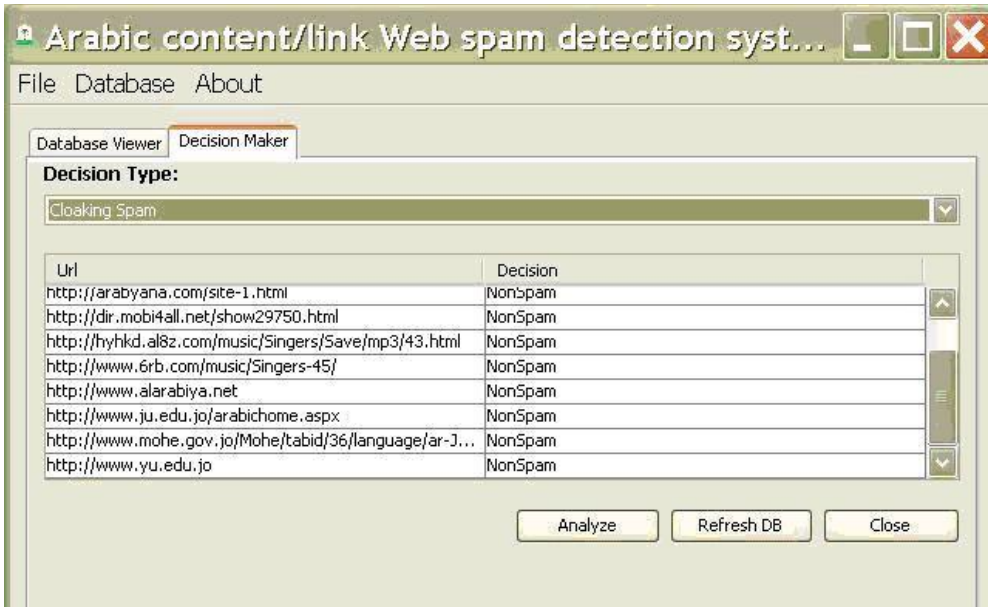

Figure 5: Evaluation process of Arabic cloaking Web spam detection system

The detailed evaluation results of Arabic Cloaking Web spam are shown in Table 4.

Table 4. Evaluation results of Arabic cloaking Web spam detectin system

Test Dataset	Accuracy	Error	True Positive Rate	False Positive Rate	Precision	Recall	F-Measure	ROC
spam	-	-	0.929	0.045	0.956	0.929	0.942	0.995
Non spam	-	-	0.995	0.071	0.928	0.955	0.941	0.995
All	94.1606%	5.8394%	0.942	0.058	0.942	0.942	0.942	0.955

7. Conclusions and Future Work

The continuous expansion of the Internet, lead to increase in the number of challenges to Web search companies to offer relevant and high quality Arabic information to its Arab users. So to accomplish their goals the Web sites owners attempt to adopt legal and illegal ways to lets their Web pages rank higher than they deserve in the SERP, to gain more users, and more revenues.

In this paper we presented novel Arabic cloaking Web spam detection system, which capable to extract large number of the cloaking features that monitor all HTML elements in both user browser version and Web crawler version. The developed system was implemented and tested on the Arabic spam dataset, and yields acceptance results that would save the time of Arabic users, efforts, and help to retrieve the relevant results that satisfy their needs.

We plan in the future to enhance the work of Web spam detection by including all the challenges of spam factors that influence the reputable Web pages and link popularities. And also detect the spam techniques in the social networks, such as: FaceBook, YouTube, Google + and Twitter. As they are attracting more and more internet users, and they are targets for spammers.

References

[1] W. Dou, K. Lim, C. Su, N. Zhou and N. Cui, "Brand Positioning Strategy Using Search Engine Marketing", MIS Quarterly, vol. 34, no. 2, **(2010)**.
[2] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval: The Concepts and Technology behind Search", Addison-Wesley Professional, **(2010),** Indianapolis, Indiana, USA.
[3] L. Becchetti, C. Castillo, D. Donato, *et al.*, "Web spam Detection: Link-based and Content-based Techniques", In The European Integrated Project Dynamically Evolving, Large Scale Information Systems (DELIS): proceedings of the final workshop, **(2008)**, pp. 99-113; Barcelona, Spain.
[4] A. Ntoulas, M. Najork, M. Manasse, *et al.*, "Detecting spam Web Pages through Content Analysis", In Proceedings of the World Wide Web Conference, Edinburgh, Scotland. **(2006)**, pp. 83–92.
[5] Z. Gyongyi, H. Garcia-Molina and J. Pedersen, "Combating Web spam with TrustRank", In Proceedings of the 30th International Conference on Very Large Databases (VLDB), **(2004)**, pp. 576-587; Toronto, Canada.
[6] L. Ermakova, "Transforming Message Detection", Young Scientists Conference in Information Retrieval, **(2011)**, pp. 15-29; Voronezh, Russian.
[7] Y. Wang, M. Ma, Y. Niu and H. Chen, "Spam Double-Funnel: Connecting Web spammers with Advertisers", International World Wide Web Conference Committee (IW3C2), **(2007)**, pp. 8–12; Banff, Alberta, Canada.
[8] G. Boone, J. Secci and L. Gallant, "Emerging Trends in Online Advertising", doxa comunicacion., vol. 5, no. 5, **(2009)**.
[9] 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, no. 1, **(2012)**.
[10] K. Narisawa, H. Bannai, K. Hatano and M. Takeda, "Unsupervised spam Detection based on String Alienness Measures", In Proceedings of the 10th international conference on Discovery science Pages (DS'07), ACM, **(2007)**, pp. 161-172; Banff, Canada.
[11] T. Jones, D. Hawking and R. Sankaranarayana, "A Framework for Measuring the Impact of Web spam", In Proceedings of the 12th Australasian Document Computing Symposium, **(2007)**, pp. 108-111, Melbourne, Australia.

[12] H. A. Wahsheh and M. N. Al-Kabi, "Detecting Arabic Web spam", The 5th International Conference on Information Technology, ICIT 2011, **(2011)**, pp. 1-8; Amman-Jordan.
[13] 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, **(2011)**.
[14] 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), **(2011)**, pp. 1-4; Amman Jordan.
[15] 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, **(2012)**.
[16] 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, **(2012)**.
[17] 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, **(2012)**, pp. 486-490; Al-Madinah Al-Munawwarah, Saudi Arabia.
[18] J. Caverlee and L. Liu, "Countering Web spam with Credibility-Based Link Analysis", In Proceedings of the annual ACM Symposium on principles of Distributed Computing, **(2007)**, pp. 157-166, Portland, Oregon, USA.
[19] Y. Chung, M. Toyoda and M. Kitsuregawa, "Identifying spam link generators for monitoring emerging web spam", In Proceedings of the 4th workshop on Information credibility (WICOW '10), **(2010)**, pp. 51-58; Raleigh, North Carolina, USA.
[20] P. Hayati, K. Chai, V. Potdar and A. Talevski, "Honeyspam 2.0: Profiling Web spambot Behavior", In Proceedings of the 12th International Conference on Principles of Practice in Multi-Agent Systems (PRIMA '09), **(2009)**, pp. 335-344, Nagoya, Japan.
[21] P. Hayati and V. Potdar, "Toward spam 2.0: An Evaluation of Web 2.0 Anti-spam Methods", 7th IEEE International Conference on Industrial Informatics (INDIN 2009), **(2009)**, pp. 875-880, Cardiff, Wales.
[22] P. Hayati, K. Chai, V. Potdar and A. Talevski, "Behavior-Based Web spambot Detection by Utilising Action Time and Action Frequency", In Proceedings of International Conference for Computational Science and its Applications (ICCSA), **(2010)**, pp. 351-360; Fukuoka, Japan.
[23] P. Hayati, K. Chai, V. Potdar and A. Talevski, "Web spambot Detection Based on Web Navigation Behavior", In Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications, **(2010)**, pp. 797-803, Perth, Australia.
[24] 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, **(2012)**, pp. 1-5.
[25] J. Lin, "Detection of cloaked Web spam by using tag-based methods", Expert Systems with Applications, vol. 36, **(2009)**.
[26] B. Wu and B. Davison, "Cloaking and Redirection: A Preliminary Study", Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web AIRWeb'05, **(2005)** 1-10; Chiba, Japan.
[27] K. Chellapilla and A. Maykov, "A taxonomy of JavaScript redirection spam", In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web (AIRWeb '07), ACM, **(2007)**, pp. 81-88; Banff, Alberta, Canada.
[28] C. Castillo, D. Donato, A. Gionis, V. Murdock and F. Silvestri, "Know your neighbors: Web spam detection using the Web topology", In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, **(2007)**, pp. 423-430, Amsterdam, Netherlands.
[29] S. Webb, J. Caverlee and C. Pu, "Characterizing Web spam Using Content and HTTP Session Analysis", In Proceedings of the Fourth Conference on Email and Anti-spam (CEAS 2007), **(2007)**, pp. 1-9; Mountain View, California, USA.
[30] Y. Tian, G. Weiss and Q. Ma, "A Semi-Supervised Approach for Web spam Detection using Combinatorial Feature-Fusion", In Proceedings of the Graph Labeling Workshop and Web spam Challenge (GRAPHLAB 2007), **(2007)**, pp. 16-23, Warsaw, Poland.
[31] K. Vangapandu, D. Brewer and K. Li, "A Study of URL Redirection Indicating spam", In Proceedings of the Fourth Conference on Email and Anti-spam (CEAS 2009), **(2009)**, pp. 1-9, California, USA.
[32] J. Abernethy, O. Chapelle and C. Castillo, "Web spam Identification Through Content and Hyperlinks", Fourth International Workshop on Adversarial Information Retrieval on the Web AIRWeb '08, **(2008)**, pp. 41-44; Beijing, China.
[33] 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, **(2009)**, pp. 21-28; Madrid, Spain.

[34] N. Spirin and J. Han, "Survey on Web spam Detection: Principles and Algorithms", SIGKDD Exploration, vol. 13, no. 2, **(2011)**.
[35] Witten and E. Frank, "Data Mining: Practica Machine Learning Tools and Techniques", Morgan Kaufmann Series in Data Management Systems, second edition, Morgan Kaufmann (MK), **(2005)**.

Authors

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

Mohammed Al-Kabi Mohammed Al-Kabi, born in Baghdad/Iraq in 1959. He obtained his Ph.D. degree in Mathematics from the University of Lodz/Poland (2001), his masters degree in Computer Science from the University of Baghdad/Iraq (1989), and his bachelor degree in statistics from the University of Baghdad/Iraq(1981). Mohammed Naji AL-Kabi is an assistant Pr ofessor in the Faculty of Sciences and IT, at Zarqa University. Prior to joining Zarqa University, he worked many years at Yarmouk University in Jordan, the Nahrain University and Mustanserya University in Iraq. AL-Kabi's research interests include Information Retrieval, Web search engines, Data Mining, Software Engineering & Natural Language Processing. He is the author of several publications on these topics. His teaching interests focus on information retrieval, Web programming, data mining, DBMS (ORACLE & MS Access).

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