I.J. Intelligent Systems and Applications, **2013, 01, 30-43**
Published Online December 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2013.01.03

A link and Content Hybrid Approach for Arabic Web Spam Detection

Heider A. Wahsheh[1], Mohammed N. Al-Kabi[2], Izzat M. Alsmadi[1]
[1]Dept. of Computer Information Systems, IT & CS Faculty, Yarmouk University, Irbid, Jordan
heiderwahsheh@yahoo.com; ialsmadi@yu.edu.jo
[2]Faculty of Sciences and IT, Zarqa University, Zarqa, Jordan
mohammedk@zpu.edu.jo

Abstract— Some Web sites developers act as spammers and try to mislead the search engines by using illegal Search Engine Optimizations (SEO) tips to increase the rank of their Web documents, to be more visible at the top 10 SERP. This is since gaining more visitors for marketing and commercial goals. This study is a continuation of a series of Arabic Web spam studies conducted by the authors, where this study is dedicated to build the first Arabic content/link Web spam detection system. This Novel system is capable to extract the set of content and link features of Web pages, in order to build the largest Arabic Web spam dataset. The constructed dataset contains three groups with the following three percentages of spam contents: 2%, 30%, and 40%. These three groups with varying percentages of spam contents were collected through the embedded crawler in the proposed system. The automated classification of spam Web pages used based on the features in the benchmark dataset. The proposed system used the rules of Decision Tree; which is considered as the best classifier to detect Arabic content/link Web spam. The proposed system helps to clean the SERP from all URLs referring to Arabic spam Web pages. It produces accuracy of 90.1099% for Arabic content-based, 93.1034% for Arabic link-based, and 89.011% in detecting both Arabic content and link Web spam, based on the collected dataset and conducted analysis.

Index Terms— Arabic Web Spam, Content-Based Detection, Link-Based Detection, Content/Link Arabic Web Spam

I. Introduction

The Internet has become the largest ever information reservoir humanity ever known. This huge reservoir of information consists of a large number of heterogeneous networks of computers, which stored a large number of various Web documents, such as audio, video, text and other interactive media features. Internet contains information in different natural languages, and characterized by the wide range of topics being presented to Internet users such as: news, sports, politics, economics, entertainments, and education.

Internet is used around the world for different purposes. Some are using it for communications, while others use it for entertainment through the use of social networks such as: Facebook, Google plus, and Twitter, or through email services, and instant messaging. The Internet is used by vast amount of users to check the latest news and weather conditions within their own countries. Some use it as an educational platform. Users usually use search engines and directories as a portal to this amazing world of information [1].

The Arab World constitutes about 5% of the world population, only 3.3% of the total number of Internet users are Arab users and the Arabic content on the Internet is less than 1% of all available online content [2, 3, 4].

The usage of Internet throughout the Arab world is witnessing a rapid increase every day, particularly in the fields of social networks, and e- commerce [3].

The statistics show that the United Arab Emirates, Bahrain, and Qatar are at the top of penetration rates list, while Iraq and Somalia are at the bottom of this list. These differences in the Internet Penetration Rate (IPR) can be explained by the regulatory, political environments, and the absence of mutual strategies to encourage the use of the Internet between the Arab countries [2, 3].

Many Arabic Web pages are characterized by having unstructured format, lacking quality of the Arabic content and containing poor information, where the poor contribution appears within blogs and forums, which constitutes around 35% of total Arabic Web content [3]. The rest of Arabic content is distributed through e-commerce, newspapers, educational Web sites, and e-government Websites. As we know the free encyclopedia (Wikipedia) allows Internet users to publish and edit different articles in more than 200 natural languages [5], the contribution of Arabs is reflected by the percentage of the Web Arabic content which does not exceed 1% in best cases [3].

Different search engines use different ranking algorithms which adopt many factors and metrics to manage the ranking process for different Web pages [6]. These factors and metrics include both content and links

features. Ranking algorithms represent a secret for different search engines; therefore these companies do not provide details about how they exactly rank the Web pages and consider these algorithms as their top secret that should not be know by other competitive search engine companies and Web spammer [7]. Web crawlers or robots constitute another part of a search engine; they are responsible for visiting different Web documents to be indexed [8].

Web spam refers to any illegal process aims to increase the rank of poor-quality Web pages and Web documents. This returns unrelated results to user's query [9].

The owners of Web sites and the Webmasters of commercial Websites use Search Engine Optimization (SEO), Search Engine Marketing (SEM), to be more visible in SERPs, and the Banner advertisements, to attract user traffic, which increase company revenues [10, 11].

SEO is based on the number of ethical methods and techniques aiming to reformat the content and material posted to the Internet. It helps the Web pages to meet the search engine requirements and gets a good rank to be considered as relevant Web pages in SERPs. SEO depends on the most essential content HTML tags such as: <title>, Anchor text, URL, Headers tags (<h1>...<h6>), , and <Meta> tags. The improvement of the content of these tags will help the Web page to rank higher within SERPs [11]. SEM techniques also called pay per click (PPC) marketing are interested in optimizing the commercial Web pages. It helps the business growth, as it suggests the most popular marketing Keywords to appear inside the most important weighting tags in the Web pages. SEM is distinguished from SEO by its adopted technique which includes both pay per click (Adwords) and SEO [10, 11].

Banner advertisements are constructed from the attractive elements like graphics, animations, flashes, sounds, and videos are used to create banners, which usually linked to the company Web site advertisements. Banner advertisement seems more advantageous than SEO and SEM, because it is based on the idea of eye catching graphics which attract more users to click on the banners, and visit the Web pages advertisements [11].

Some of Web site owners act as spammers or try to hire Web spammers, using the illegal SEO, SEM, and Banner advertisements methods and techniques in a complete or partial way to increase the rank of their Web pages. These methods define the term "Web spam" which fill the Internet with Web pages that deceive the search engines and take higher ranks than what they really deserve [9].

There is a lack in dataset collections related to Arabic content/link Web spam, and this is considered the main problem affecting the research in this field. In addition, the researches so far are few. So this paper adopts an enhancement to the previous studies of content-based/link-based Arabic Web spam detection [12-19]. We collected a larger Arabic Web spam dataset to improve the Arabic Web spam features mentioned in the previous studies, and built a content/link Arabic Web spam detection system.

In this paper we collected a larger Arabic content/link based spam dataset than those collected in the previous Arabic Web spam studies. We adopted advanced content/link based features. The extracted features will be fed into classification algorithms, such as: Decision Tree, Logistic, and *K*-Nearest Neighbor (*K-NN*). The results of the classification algorithms are compared, and the best algorithm identified. Forwards, the rules of the best algorithm are implemented to build Arabic Web spam detection system.

The next sections of this paper organized as follows: Section two presents' related studies to Web spam detection, section three shows the proposed methodology, section four presents implementation and experimental results, section five described the evaluation results of our proposed system. Last but not least section sex presents the conclusions and future work.

II. Related Work

Many studies were conducted to explore different techniques to detect Web spam. This section presents these techniques and categorizes them into four sections. The first section presents non Arabic content-based spam studies, the second section presents non Arabic link-based spam studies, the third section dedicated to the non Arabic content/link spam studies, and the fourth section presents the earlier Arabic content-based Web spam studies.

2.1 Non Arabic content-based Web spam detection

In their study [20] use a various content-based features extracted from a real dataset of spam Web pages. They used a number of heuristic methods for detecting content-based spam, and achieving high accuracy of detection using C4.5 classifier, which correctly identifies 86.2% of spam Web page within the dataset.

[21] produce a novel content-based trust model for Web spam detection, according to two real datasets one is in English and the other one in Chinese languages. The results showed an enhancement on Web spam detection using SVM which yielded an accuracy of 90.13%.

In their study [22] propose a set of content-based features in which the occurrence of keywords play the main role in identifying the Web pages as spam or high value advertisement Web page. The experiment tests

applied on the public known WEBSPAM-UK2006 dataset, and the results improved more than 3%.

The Term Frequency - Inverse Document Frequency (*TF-IDF*) is a weighted schema which shows the importance of the words in the document [23]. The value of *TF-IDF* of each term is dependent on the frequency of that term beside the number of documents which has that term, and the occurrence of the terms inside the Web pages [23]. The terms which appears in special positions in the Web page such as: the <body> tag, Anchor text, URL, Headers (<h1>...<h6> tags), <meta> tags, and within the Web page <title> present more important than the other terms in the rest of Web page positions [24].

2.2 Non Arabic Link-based Web Spam Detection

The study of [25] presents an algorithm dedicated to the link spam detection, called R-spamRank. This algorithm produces an automated selection of spam Web pages especially those appears in the link farms. The authors manually collected a small spam dataset which considered as seeds for the evaluation process. They assigned spam values to the Web pages, and selected semi-automatically the most likelihood spam Web pages. The results of this algorithm yielded an accuracy of 91.1% in detecting Web spam.

[26] Study special technique of link-based Web spam called hijacked links spam; which is based on bringing the rank scores from normal Web pages to the target Web spam pages. [26] propose an algorithm for link hijacking detection, which is based on analyzing the features of the link structure which is neighbor to the hijacked Web sites. The results showed improvement in the accuracy of detecting the hijacked Web sites, where around 25% better relative to the other previous approaches.

[27] Build a link spam dataset which contains over 235,000 links of English Wikipedia, with extracted 40 features, by using Wiki metadata, landing site analysis, and external data sources. The conducted results showed enhancement in link-based Web spam detection.

[28] Continue to work on the semi-supervised learning algorithm, by proposing a novel algorithm; called Harmonic Functions based Semi-Supervised Learning (HFSSL), where the labeled and unlabeled Web pages given weights based on the similarity in weighted Web graph. The results showed enhancement in the Web spam detection.

2.3 Non Arabic content/link Web spam detection

A Quantitative Study of forum spamming which uses a context-based and reported by [29]. The importance of the spam forums and splogs is due to three main perspectives: search users, spammers, and forum sites. The study of [29] focus on the content-based and cloaking spam, and showed that the spam forums were used extensively. The spam forums supported by popular forum programs (which able the spam forums to occupy the top 20 search results for most popular keywords). The spam comments also used to increase the traffic on the honey spam forums. The results of splogs showed that more than half blogs are spam. The researchers proposed context-based analysis; based on the cloaking analysis, to automatically detect spam.

[30] study several content-based and link-based Web spam techniques such as: keywords stuffing, links stuffing, redirection pages, duplicated content, and hiding text. Two well known datasets were used in the experiments UK2002; which contains 18.5 million Web pages and WEBSPAM-UK2006 which consists of 77.9 million Web pages. Their experiments results revealed that the adopted techniques were able to detect around 88% of spam Web pages.

[31] study has focused on different methods for Web spam detection. A novel approach used machine learning to build Web spam detection tool. The UCINET software and SVM classifier were used to identify the spam Web pages, based on many proposed features such as: degree of centrality, links betweenness and Eigen vector value of the link, which identify the quantitative and qualitative link farm properties. Their proposed approach used the WordNet database through the semantic analyzer, and obtained useful information that successfully discovered the spam Web pages.

In the study of [32] the researchers develop a new system called Spamizer which able to detect the spam Web pages using content- based and link-based features. The Spamizer analyzed several available link spam algorithms, such as: Relative spam Mass Estimation, Trustrank, Anti-Trustrank, Propagating Trust, and Distrust Scores and Reverse spam Rank. The experiments used the public known spam dataset WEBSPAM-UK2007, and they found that integrating the spamicity scores that generate from each algorithm increase the predictability for the spam and non spam Web pages.

2.4 Arabic content/link based Web spam detection

In their study, the authors in [12] conducted a series of studies dedicated to Arabic Web spam problem. The authors in [12] have manually collected a small Arabic Web spam dataset; containing around 400 Arabic content-based spam Web pages. Three classifiers were tested; Decision Tree, Naïve Bayes, and *K*-Nearest Neighbour (*K-NN*). The results showed that the *K-NN* yielded a better accuracy than the other two classifiers in detecting Arabic Web spam pages. The study of [13] follows the [12] and proposed new content-based features to improve the Arabic Web spam detection. Their study applied three classifiers (Decision Tree, Naïve Bayes, and LogitBoost), and their results showed that the Decision Tree classifier achieved the best results.

In [14] the authors have integrated the two previous studies [12, 13], and propose a set of new content-based

features, and used a larger spam dataset than [12]. Three classification algorithms (Decision Tree, LogitBoost, and SVM) were used to detect Arabic Web spam. The results confirmed the superiority of the Decision Tree as the best classifier with an accuracy of 99.3462% to detect Arabic Web spam.

The study of [15] analyzes the behaviors of the spammers to create spammed Arabic Web pages. They computed the weights of the most ten popular Arabic words used in the content of the HTML tags, which used in the Arabic queries. The results present special key stuffing techniques used in the Arabic spammed Web pages. The conducted tests used the Decision Tree classifier to evaluate the spammer's behavior, and achieved 90% accuracy to detect Arabic Web spam.

The study of [16] improves their previous studies on the content-based Arabic Web spam. They used a large Arabic content-based spam dataset which contains 15,000 Web pages, that were collected by a special crawler. These Web pages were identified manually as spam or non spam. They applied four different classification algorithms (Naïve Bayes, Decision Tree, SVM, and *K-NN*) on the groups of the datasets, where the spam percentages were: 1%, 15%, and 50%. The results also revealed that the Decision Tree was the best classifier with 99.96% accuracy.

Machine learning is used to identify spam Web pages. [17] conducted a study based on the machine learning algorithm to identify the content-based Arabic spam Web pages. The spam dataset was collected from three resources: the first is Extended-Arabic-2011 Web spam dataset, and the second is UK-2011 spam dataset where they were built by [17]. The third is a portion of the WEBSPAM-UK2007 spam dataset. Experiments were based on two algorithms (Naïve Bayes, and Decision Tree). The conducted tests of the proposed features show high accuracy results with Decision Tree which is better than Naïve Bayes in detecting Arabic spam pages, and yields sufficiently good results in detecting non Arabic Web spam.

All the previous Arabic Web spam studies [12-17] tried to identify the best classification algorithm for the content-based Arabic Web spam detection, which almost unanimously indicate the Decision Tree classifier is the best. Therefore [18] based on the 15,000 Arabic spam Web pages, enhanced more content-based features, and built the novel Arabic Web spam detection system using the rules of Decision Tree classification algorithm. The experiment results presented an accuracy of 83% using the proposed system.

In an attempt to solve the problem of the Arabic link-based Web spam, [19] studied the link-based spamming technique which is used by Arabic Web spammers. [19] present that the spammers used the link-based spam techniques in the Arabic Web pages. The first Arabic link-based spam Web pages dataset was built by them. Many link-based features were extracted, and two classifiers (Decision Tree, and Naïve Bayes) were applied to evaluate the Arabic link-based Web spam. The conducted experiment show that spammers use a link spam farms technique between Arabic spam Web pages. The results of Decision Tree yield an accuracy of 91.4706% to detect link-based spam Web pages.

III. Research Methodology

In this section we present the research methodology that we used to build Arabic content/link Web spam detection system. The methodology of the Arabic Web spam detection system includes the following seven main steps:

1. Develop an embedded Web crawler; which is an automated tool, embedded in our new Arabic Web spam detection system. This crawler download the Web pages, parsed all the hyperlinks, and the content of each Web page.
2. Build a larger dataset of Arabic content/link spam Web pages relative to those built in the previous studies. The first part is called training dataset used to build Arabic content/link Web spam detection system, and contains around 18,000 Web pages. While the second part is called test dataset, contains around 5,000 Web page, and used to evaluate Arabic content/link Web spam detection system. The new dataset extended the last datasets used by [12-19], using the enhanced Web crawler.
3. Develop a Web page analyzer to extract larger number of features relative to those used in the previous studies.
4. Use the three classification algorithms Decision Tree, Logistic Regression, and *K-NN* which are supported by Weka.
5. Compare the classification results of Decision Tree, Logistic Regression, and *K-NN* algorithms to identify the best classifier to detect Arabic spam Web pages.
6. Extract the rules of the best classification algorithm using the training dataset, to develop the decision maker as a final part of our Arabic content/link Web spam detection system.
7. Evaluate the Arabic content/link Web spam detection system, using the test dataset that contains around 5,000 Web pages including spam and non spam.

Figure 1 presents a summarization of the research methodology procedures being followed in this study.

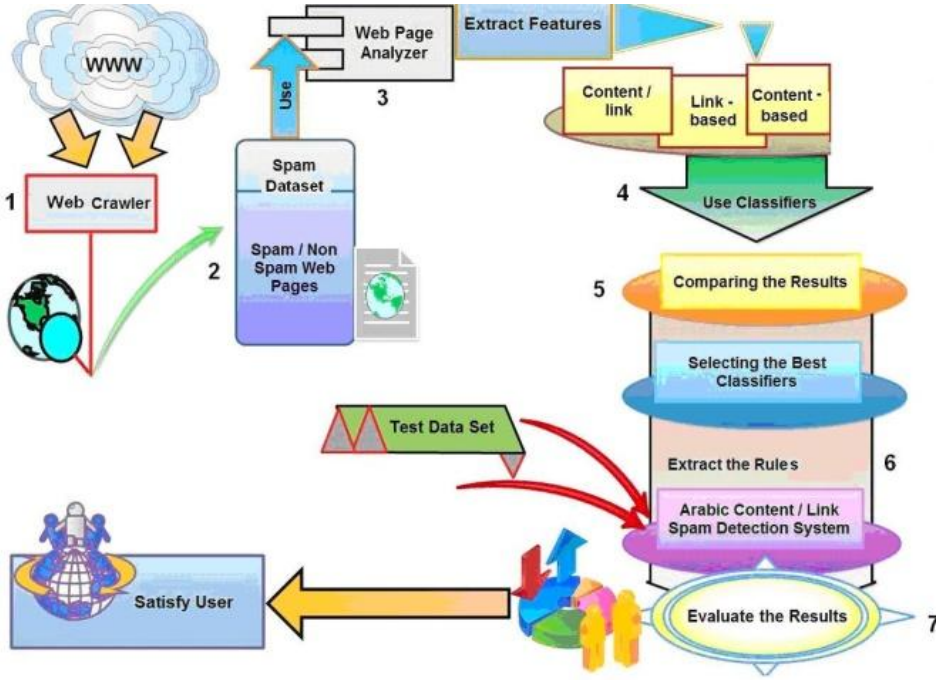

Fig. 1: Methodology procedures

3.1 Develop an Embedded Web Crawler

Web crawlers are also known as spiders, or Web agents, is a type of software agent, which automate the traversing, fetching, sorting, and clustering Web pages, creates a copy of all Web pages before indexing. Crawlers traverse the Web by starting from a random Web page and continue by the following links to other Web pages. Sometimes crawlers need to exchange the information with other crawlers in order to notify their peers about sites with rich semantic content [8]. So it appears as the main process in any search engine.

In this study we develop an embedded crawler in Arabic web spam detection system using Java programming language. This crawler enables our system to browse Arabic Web sites within the World Wide Web (WWW), and fetch different Arabic Web pages. Therefore this system is capable to download the Web pages, and automatically parsing the elements, hyperlinks, and the content of the downloaded Web pages.

3.2 Build an Arabic Web Spam Dataset

Arabic Web spam detection studies lacks generally to a large collection of Arabic spam Web pages, and this is one of the main challenges to the researchers in this field. In this paper we have built a large Arabic Web spam dataset containing around 23,000 Web pages, where 18,000 Web pages of the total Web spam dataset were used as training dataset which extended the datasets mentioned in [12-19]. The rest of the Web spam dataset consists of 5,000 Web pages that were used as test dataset. The new dataset improves both the number of Arabic spam pages and their features as shown in the next sections.

The Web pages in the Arabic Web spam dataset were divided into two types: spam, and non spam Web pages. We split the spam training dataset into many groups based on the accuracy percentages of Arabic Web spam detection. Some groups got a close accuracy values to each other; therefore we select the best percentage from those close accuracy percentages. The best three groups with different spam percentages were (2%, 30%, and 40%) of the dataset. Table 1 shows the Arabic spam dataset groups taxonomy.

Table 1: New Arabic spam dataset groups taxonomy

Close Percentages values Groups	Best Percentages values spam Group	Number of spam Web page	Number of Non spam Web Page
1%-15%	2%	460	22540
16%-32%	30%	6900	16100
33%-50%	40%	9200	13800

We have manually labeled the Web pages as either spam or non-spam pages based on the authors' judgments, previous Arabic Web spam studies, and based on similar types of Non Arabic spam Web pages, depending on the spam content-based features for some Web pages, mislead links, and the reputations of the Web pages.

The spam Web pages consist of Arabic content-based blogs, forums, some of marketing Web pages, and advertising Web pages, which tries to increase their possibility to appear at the top of SERP.

The non spam Arabic Web pages can easily be found within the Web pages of universities and educational (.edu), ministries and governmental institutions (.gov), news sites (.com, .net) well known business companies (.com), and satellite channel (.tv).

3.3 Develop Web Page Analyzer

We developed a Web page analyzer capable to extract the previous proposed features by [12-19], and proposed new content ad link features.

3.3.1 Content-based Features

Arabic Web spammers used Keyword stuffing to increase the rank of their spam Web pages. The keyword stuffing spamming technique is based mainly on the duplication of some words in the main HTML elements. Arabic Web pages spammers used a unique keyword stuffing technique which is based on duplicating meaningless English words. Unique keyword stuffing technique is based on the relation between the meaningless English words and their corresponding Arabic letters that lie on the Arabic/English key board [12-19]. This spam behavior leads to increase the rank of Arabic spammed Web pages, and from our point of view lead to deterioration of the quality of the Arabic content [12-19].

In [15] the authors found that the spammers used meaningless English words in Arabic spammed Web pages. The Latin letters of those meaningless English words lie on the same Arabic/English keyboard keys. Therefore the lengths of these meaningless English words always equal the lengths of Arabic words. Top ten Arabic words used by [15] are not enough to detect the meaningless English words spammed technique. Thus, to address all the different topics that spammers might seek to use it in their techniques. Our Web page analyzer convert every English word to its corresponding meaningful/meaningless Arabic word, by converting each Latin letter of the English word to its Arabic letter sharing the same key on the Arabic/English keyboard. Then the analyzer use the database which contains Arabic word list which developed by [33]. The database of Arabic word list contains nine million Arabic words. So the analyzer check the availability of every converted Arabic word in the database, to determine whether it is meaningful or not. Our analyzer considered the converted Arabic word a spammed word if it is found in the database. Spammed Arabic words are generally meaningless.

Figure 2 shows an example of Arabic spam Web page using meaningless English words Keyword stuffing technique. The Arabic words not include in the top ten Arabic Keywords used in the search engines, but the spammers used them to increase the rank of the Arabic Web page to increase the *TF-IDF*, and it is reducing the quality of the Arabic Web pages.


فلاش العاب بنات العاب اكشن موقع العاب العاب اطفال العاب فقط للبنات
العاب للبنات العاب باربي العاب للبنات فقط العاب فلاش العاب بنات العاب
جديدة العاب الذاكرة العاب اطفال العاب اكشن العاب ذكاء العاب بازل و
متاهات العاب العاب بنات العاب فلاش العاب الفلاش العاب بنات، ألعاب بنات،
العاب تلبيس بنات، مركز العاب، العاب مكياج بنات، العاب اكسسوارات بنات،
العاب ازياء بنات، العاب فتيات، العاب تلبيس، العاب فلاش، العاب مغامرات،
العاب فلاشية، العاب اطفال، العاب فضائية، العاب قتال، العاب منوعة العاب
اثاره ومغامرات العاب اثاره العاب مغامرات ازياء بنات مكياج واكسسوارات
العاب رياضية العاب الغاز العاب تفكير العاب ذكاء العاب ملابس بنات العاب
ملابس وازياء العاب ازياء العاب اوراق العاب فضائية العاب اطفال العاب
للاطفال العاب اكشن العاب مغامرات العاب قويه العاب جديده العاب مغامرات
العاب طبيب العاب طبخ العاب العاب فلاش العاب بنات العاب فتيات العاب
للبنات العاب للفتيات العاب باربي ألعاب منوعة العاب اطفال العاب للطفل
العاب كرة قدم و رياضه و لعبة سباق سيارات
hguhf ugn ;dt;, hguhf ugn ;dt;, hguhf hgjsgdm ,hgjvtdi hguhf lk,ui
hguhf lqp;m hguhf h;ak hguhf lyhlvhj hguhf fkhj hguhf lh;dh[hguhf
h'thg hguhf pvfdm hguhf vdhqm hguhf sdhvhj hguhf svum lghp/m
hguhf ,vr hguhf lyhlvm hguhf `;hx hguhf fhvfd hguhf jgfds hguhf
lyhlvhj hguhf []d]i hguhf []d]i hguhf l[hkdm hguhf tgha hguhf fhvfd
hguhf 'fo Hguhf hgjvtdi , hgjsgdi hguhf jsgdm hguhf ,jsgdm hguhf
fkhj hguhf l;dh[hguhf fhvfd hguhf ggwyhv hguhf h'thg l;dh[guf fhvfd
hguhf fhvfd hguhf pg,i hguhf [ldgm hguhf l[hkdm hguhf l[hkdm hguhf
tgha hguhf tgha l[hkdi l[l,um hguhf tgha jk.dg hguhf tgha hguhf tgha
lhvd hguhf tgha vdhqdm hguhf tgha hguhf jnlda l,u tgha hguhf

Fig. 2: Arabic spam Web page using meaningless English words Keyword stuffing technique

The Web page analyzer computes the content-based features which mentioned in the Arabic literature [12-19]. The Web page analyzer extracts the following six content-based features categories:

1. The first category contains one content-based feature which checks if we have a number of meaningless (Arabic/English) in the HTML elements which considered as duplication or as a keyword stuffing technique.
2. The second category check if we face a keyword stuffing technique, where the following two parameters are computed:
   2.1 Compute the difference between the total number of Arabic/English words inside the <body> element, and the total number of unique Arabic/English words inside the <body> element. If the results greater than or equal two third of the total number of Arabic/English words inside <body>. This means that we have spam behavior.
   2.2 Compute the difference between the total number of Arabic/English words inside a specific Web page (in all HTML tags in the Web page), and the total number of unique Arabic/English words inside a specific Web page. Our system indicates that a specific Web page is a suspected spam Web page, if the total number of unique Arabic/English words inside it is greater than or equal two third of the total number of Arabic/English words inside the Web page under consideration.
   2.3 The third category contains a number of content-based features such as: the number of Arabic popular words, the size of compression ratio (in Kilo bytes), page size (in Kilo bytes), the maximum Arabic/English word length inside (<body>, or a specific Web page), the size of hidden text (in Kilo bytes), and the total number of images.
3. The fourth category contains content-based features as follows: the maximum Arabic word length inside (<body>, or a specific Web page), the average lengths of Arabic/English words inside the (<body>, or a specific Web page), the average lengths of Arabic words inside the (<body>, or a specific Web page), maximum Arabic/English word length inside a specific Web page, total number of characters of the symbols in all <Meta>, average length of English words inside a specific Web page, average length of Symbol words inside a specific Web page, number of unique Symbol words inside a specific Web page, and number of English popular words.
4. The fifth category contains other influential content-based features as follows: number of images-links inside a specific Web page, maximum symbol word length inside a specific Web page, number of Arabic/English words inside <title>, compressed files inside a specific Web page, number of English characters inside <Meta>, number of English characters inside a specific Web page, number of characters of the symbols in all <Meta>, and the number of unique symbol words inside a specific Web page.
5. The sixth category contains the last influential content-based features such as: number of <Meta> elements inside a specific Web page, number of characters within the URL, number of Arabic/English characters inside a specific Web page, average lengths Arabic/English words inside a specific Web page, average lengths of Arabic words inside a specific Web page, number of <Meta> elements inside a specific Web page, number of Arabic words in each <Meta> elements, number of Arabic characters inside <Meta>, number of Arabic/English characters inside <Meta>, number of Arabic/English characters inside a specific Web page.

3.3.2 Link based features

We have many types of link-based Web spam, the spammers try to create the link structure between their spam Web pages such as the following:

- Spam link farm. The spam link farm as we mentioned in the literature create heavily connected Web pages, in order to mislead search engines, where this technique is based on manipulating internal links and external links of a number of connected Web pages forming the link farm [34].
- Using the expired domains. Spammers take the benefits of expired domains by inserting their spammed Web pages in. Also the spammers try to trick the users, when they used the names which are similar to the popular trusted and reputable domains names [9].
- Link spam comments in the blogs: The spammers may post the links to spam Web pages as a comment to the blogs. So the spam comments will increase the traffic on the honey spam blogs and forums [29].

[19] exhibits that the spammers used two types of link-based techniques in Arabic Web pages. The first type is based on using the spam link farms, which manipulate internal and external links in the Web pages, in order to increase the rank of these spam Web pages. While the second type is based on using the expired domains which try to trick the users, when the spammers used the names which are similar to the popular trusted and reputable domains.

Figure 3 presents an example of Arabic link-based spam Web page which is full of advertisements. This type of link-based Web spam is called scraper Web

page that does not contain any real content related to the Website, where the links redirect users to other Web sites.

Fig. 3: Example of Arabic link-based spam Web page

The Web page analyzer extracts the following link-based features:

1. The number of external links within the page under consideration.
2. The number of internal links within the page under consideration.
3. The total number of links (the internal and external) within the page under consideration.
4. The URL length (total number of characters in URL).
5. The total number of broken links. It is also known as dead link within the page under consideration. The broken links are called broken, due they are no longer point to non spam Web pages, so they decrease the rank of a Web page [35].
6. The total number of redirected links within the page under consideration.
7. The total number of empty link text (links without anchor text) within the page under consideration.
8. The total number of empty links (anchor text without links) within the page under consideration.

IV. Implementations and Experimental Results

In this section we extract the content/link features using our Web page Analyzer, apply three classifiers (Logistic Regression, Decision Tree, and *K-NN*) on three groups within one dataset with various percentages of spam Web pages (2%, 30%, and 40%). Afterward we compare the results to identify the best classifier, capable to detect Arabic content/link Web spam. Finally extract the rules of the best classifier to build the Arabic content/link Web spam detection system using Java programming language.

4.1 Arabic Content/Link Web Spam Features Extraction

We extract the content/link features using our Web page Analyzer. These features are mentioned in section 3.3.

4.2 Apply Classification Algorithms

Weka is one of the most popular data mining tools. It provides us with a number of classification algorithms such as: Logistic Regression, *K*-Nearest Neighbor (*K-NN*), and Decision Tree. These three classification algorithms are used in this study to detect if the Web page is either a spam or non-spam.

4.2.1 Logistic Regression Algorithm

Logistic Regression is one of the approaches used in regression analysis. It is widely used statistical modeling technique for predicting the outcome of categorized variables depending on the predictor variables. Binomial and multinomial regressions are two models used in logistic regression. The binomial (binary) can observe the outcome with two possible types as a (0, or 1) which expressed the straightforward interpretation, while the multinomial regression indicate that the outcomes can have more than two possible types [36].

4.2.2 K-Nearest Neighbour (K-NN) Algorithm

K-Nearest Neighbor also known as IBK in Weka. It is considered as the simplest machine learning algorithms, and it is one of the lazy learning types. The classification decision based on the closest training objects values *K*, which starts from 1, and indicates to the space of the neighborhoods around the test pattern [37].

4.2.3 Decision Tree Algorithm

The Decision Tree is one of the common classification techniques available in Weka. It is presents as a graph of decisions which consist of root node and many leaves nodes. The decision is based on the result of comparison between the values of the features and values stored on different nodes of the tree paths [38].

Decision Tree is a high speed and powerful way to express the tree structure. It is widely used in research studies to identify the decision strategy for specific goals [39].

4.2.4 Arabic Web Spam Classification Results

Applying the Logistic Regression algorithm on the three spam percentage groups for content-based, link-based, and hybrid detection approach yields different acceptance accuracies, and the 2% spam group yields the best accuracy results. The results of the accuracies are shown in the Table 2.

Table 2: Logistic Regression results

Spam Percentage Group	Accuracy (Content)	Accuracy (Link)	Accuracy (Content/Link)
2% spam Group.	98.0411%	84.7924%	75.5894%
30% spam Group.	82.3529%	79.3644%	70.2168%
40% spam Group.	95.8333%	76.5988%	67.3343%

Table 2 indicates that different percentages of spam with three different dataset groups have a significant impact on the accuracy of the Logistic Regression classifier results.

Applying the second classifier (*K-NN*) on the three spam percentage groups yields better results than the results of Logistic Regression. 2% spam group still yields the best accuracy results. The results of the accuracies are shown in Table 3.

Table 3: *K-NN* results

Spam Percentage Group	Accuracy (Content)	Accuracy (Link)	Accuracy (Content/Link)
2% spam Group.	99.7083	98.7174%	98.3437%
30% spam Group.	85.2941%	99.7008%	96.1014%
40% spam Group.	95.8333%	98.7664%	89.4434%

Finally we applied the Decision Tree classifier on the three spam percentage groups. Table 4 presents the detailed results of the accuracies.

Table 4: Decision Tree results

Spam Percentage Group	Accuracy (Content)	Accuracy (Link)	Accuracy (Content/Link)
2% spam Group.	99.7611%	99.8174%	98.2422%
30% spam Group.	88.1188%	99.7041%	97.0891%
40% spam Group.	96.875%	99.6647%	96.7218%

Table 4 shows that the Decision Tree classifier is the best in detecting all Arabic Web spam types, within the highest results using 2% spam dataset.

4.3 Rules Extraction

In section 4.2, we found that Decision Tree with (2%) spam percentage is the best to detect the Arabic Web spam types; content-based, link-based, and content/link.

We extract the rules of the Decision Tree for every spam type then we use the Java programming language to build the Arabic content/link Web spam detection system.

4.3.1 Arabic content Web spam detection System

The Arabic content-based Web spam detection system based on the previous content-based categories

which mentioned in section 3.3 (Develop Web page analyzer) which yielded the best results with Decision Tree classifier. We extracted the rules of 2% spam percentage group to build Arabic content-based Web spam detection system.

Figure 4 shows Arabic content-based Web spam detection algorithm. This algorithm used the six content-based categories (mentioned in 3.3.1).

Algorithm	**Arabic Content-based Web spam Detection System**.
Input:	**List of URLs in (ContentbasedURl.txt), or list of URLs the database stored in the system.**
Output:	**Table of the URLs with the decision as a (spam/ Non spam).**

BEGIN
Open ContentbasedURl.txt or Database
While Not EOF (ContentbasedURl.txt)
Read the URL of Web page
Download a Web page.
Call content-based Web spam categories.
Apply rules of Decision Tree algorithm.
Make a decision of non spam/ or true percentage of spam.
END WHILE
END

Fig. 4: Arabic content-based Web spam detection algorithm

4.3.2 Arabic link Web spam detection System

Depending on the rules which were extracted from the Decision Tree when applied on (2%) spam percentage group, we built the Arabic link-based Web spam detection system. Figure 5 presents the algorithm of Arabic link-based Web spam detection system.

Algorithm	**Arabic Link-based Web spam detection system.**
Input:	**List of URLs in (LinkbasedURl.txt), or list of URLs the database stored in the system.**
Output:	**Table of the URLs with the decision as a (spam/ Non spam).**

BEGIN
Open LinkbasedURl.txt or Database
While Not EOF (LinkbasedURl.txt)
Read the URL of Web page
Download a Web page.
Call link-based Web spam detection features.
Apply rules of Decision Tree algorithm.
Make a decision of non spam/ or true percentage of spam.
END WHILE
END

Fig. 5: Arabic link-based Web spam detection Algorithm

4.3.3 Arabic content/link Web Spam Detection System

Using the rules which extracted from the Decision Tree applied on a (2%) content-based and link-based spam percentage dataset. The two algorithms are merged and built the Arabic content/link Web spam detection system. Figure 6 shows the algorithm of Arabic content/link-based Web spam detection system.

Algorithm	**Arabic content/link-based Web spam detection system.**
Input:	**List of URLs in (ContentLinkbasedURl.txt), or list of URLs the database stored in the system.**
Output:	**Table of the URLs with the decision as a (spam/ Non spam).**

BEGIN
Open ContentLinkbasedURl.txt or Database
While Not EOF (LinkbasedURl.txt)
Read the URL of Web page
Call content-based Web spam detection algorithm.
Call link-based Web spam detection algorithm.
Make a decision of non spam/ or true percentage of spam.
END WHILE
END

Fig. 6: Arabic content/link-based Web spam detection Algorithm

V. Evaluation Results

In this section we evaluated all types of our Arabic content/link Web spam detection system, using Decision Tree classifier.

In this section we evaluated the capability of our built system to detect Arabic content and link-based Web spam, using the test dataset. This dataset contains 5,000 Arabic spam Web pages which are labeled manually by human experts as either spam or non-spam. Afterward we started the evaluation process by testing the accuracy of the built system to identify the Arabic spam Web pages in the test dataset. The classification results of our system are tested by Weka Decision Tree classifier. This provides the detailed evaluation results that showed the effectiveness of our Arabic content/link Web spam detection system.

5.1 Evaluating Arabic Content-Based Web Spam Detection System

We used 5,000 spam and non spam Web pages as test dataset to evaluate the Arabic content-based Web spam detection system. The results of testing our Arabic content-based Web spam detection system yields an accuracy of 90.1099% in detecting content-based Web spam.

Table 5 presents the detailed evaluation results' of our Arabic content-based Web spam detection system, with accuracy, error, True Positive Rate (TPR), False Positive Rate (FPR), Precision (P), Recall (R), F-Measure, and Receiver Operating Characteristic (ROC).

Table 5: Evaluation results of Arabic content-based Web spam detection system

Test Dataset	Accuracy	Error	TPR	FPR	P	R	F-Measure	ROC
Spam	-	-	0.97	0.52	0.91	0.97	0.94	0.88
Non spam	-	-	0.47	0.02	0.76	0.47	0.58	0.88
All	90.1%	9.8%	-	-	-	-	-	-

5.2 Evaluating Arabic link Web spam Detection System

We used the same test dataset which consists of 5,000 Arabic spammed Web pages to evaluate the Arabic link-based Web spam detection system. Arabic link-based Web spam detection system yields 93.1034% accuracy in detecting link-based Web spam.

Table 6 shows the detailed evaluation results' of our Arabic link-based Web spam detection system.

Table 6: Evaluation Results of Arabic link-based Web spam Detection System

Test Dataset	Accuracy	Error	TPR	FPR	P	R	F- Measure	ROC
Spam	-	-	0.66	0.05	0.4	0.6	0.5	0.879
Non spam	-	-	0.94	0.33	0.9	0.9	0.96	0.879
All	93.1%	6.8%	-	-	-	-	-	-

5.3 Evaluating Arabic content/link Web spam detection system

Finally we used the same test dataset which consists of 5,000 Arabic spammed Web pages to evaluate our Arabic content/link Web spam Detection System. This test yields 89.011% accuracy in detecting content/link Web spam. The full results are shown in Table 7.

Table 7: Evaluation Arabic content/link Web spam results

Test Dataset	Accuracy	Error	TPR	FPR	P	R	F- Measure	ROC
Spam	-	-	0.96	0.55	0.9	0.9	0.938	0.9
Non spam	-	-	0.45	0.03	0.6	0.4	0.544	0.9
All	89.01%	10.9%	-	-	-	-	-	-

5.4 Comparison between all types of Arabic content/link Web spam detection system

From the above subsections, we found that the Arabic link-based detection system yields more accurate results than the other Arabic Web spam types.

Table 8 shows the comparisons of the Accuracy values between all types of spam in our Arabic content/link Web spam Detection System.

Table 8: Comparison between the accuracy values for all spam types

Test Dataset	TPR	FPR	P	R	F-Measure
Content-based	0.901	0.452	0.893	0.901	0.891
Link-based	0.931	0.139	0.951	0.931	0.939
Content/link	0.89	0.47	0.87	0.89	0.88

Table 8 shows the superiority of the Arabic link-based Web spam detection system relative to others. Followed by link-based, then content-based, and the content/link Arabic Web spam detection system respectively.

We have several performance measurements to evaluate the results of this paper, such as:

1. Kappa statistic (KS): Is the statistical measure that proportionate the reduction in errors compared to the errors of a completely classification random [39].
2. Mean Absolute Error (MAE): The mean absolute error measures how the predictions are close to the actual outcomes [39].
3. Root Mean Squared Error (RMSE): Is a measure of the differences between estimated values and actual values. It is related to the error variance or standard deviation. If RMSE is closer to zero, the prediction is considered good [16].
4. Root Absolute Error (RAE). This is the error prediction which presents a percentage error of a simple prediction model [16].
5. Root Relative Squared Error (RRSE): It is relative to what it would have been if a simple predictor had been used. It is obtained by taking the square root of the Relative squared error [16].

Table 9 presents the comparisons of the performance measurements for all Arabic Web spam types.

Table 9: Performance measurements for all Arabic Web spam types

Types of Arabic Web spam	KS	MAE	RMSE	RAE	RRSE
Content-based	0.53	0.17	0.28	62.85%	80.16%
Link-based	0.46	0.09	0.25	72.6%	115.24%
Content/link	0.48	0.16	0.27	58.43%	78.16%

Table 9 presents clearly the effectiveness of our Arabic content/link Web spam Detection System.

VI. Conclusions and Future Work

Web spamming is defined as any illegal manipulation that violate the SEO tips on the content, link structure, or some other features of the Web documents to mislead the ranking algorithms of search engines to be at the top 10 of SERP, or gain the highest possible rank for their Web pages. The spammers used the spamming techniques in Arabic Web pages, which usually presented with bad quality information.

The main goal of this study is to solve the Arabic Web spam detection problem. We discussed the relation between the Arabic Web spam types. In this paper large Arabic content/link based spam dataset relative to those used in our previous studies was built and used. This large dataset contains 23,000 Arabic spam Web pages which were collected through an enhanced embedded Web crawler. The spam dataset is divided into two parts: training dataset is used to build the proposed system and test dataset is used to evaluate the proposed system.

The extracted content/link features used by three classification algorithms to identify the best algorithm to detect the Arabic content/link Web spam. The rules of Decision Tree were extracted with 2% percentage group spam dataset, which is considered as the best algorithm to detect the Arabic content/link Web spam. Then we evaluated the Arabic content/link Web spam detection system, using test dataset, and we gained good results for Arabic content/link Web spam.

We plan to extend this work in the future to study and investigate the detection of the malicious links in Arabic spammed Web pages. Malicious links usually combines between Web spam techniques and Web security issues particularly malware types (Worms and Viruses).

References

[1] W. Alrawabdeh. 2009. Internet and the Arab World: Understanding the Key Issues and Overcoming the Barriers. The International Arab Journal of Information Technology. v6, n1, 2009, pp. 27-33.

[2] Internet World Stats, 2012. Arabic Speaking Internet Users Statistics. Retrieved February, 24, 2012 from the World Wide Web: http://www.Internetworldstats.com/stats19.htm

[3] A. Tarabaouni. MENA Online Advertising Industry. Retrieved October, 28, 2011 from the World Wide Web: http://www.slideshare.net/aitmit/mena-online-advertising-industry

[4] Internet World Stats, 2012. Internet world users by languages top 10 languages. Retrieved February, 24, 2012 from the World Wide Web: http://www.internetworldstats.com/stats7.htm

[5] R. Almeida, B. Mozafar, J. Ch. On the Evolution of Wikipedia. In Proceedings of the International Conference on Weblogs and Social Media. Boulder, Colorado, USA, (2007), pp. 1-8.

[6] M. Selvan, A. C. Sekar, A.P. Dharshini. Survey on Web Page Ranking Algorithms. International Journal of Computer Applications. v41, n19, 2012, pp.1-7.

[7] M. Bendersk, W. Crof, Y. Dia. Quality-Biased Ranking of Web Documents. In Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM 11), Hong Kong, China, (2011), pp.1-10.

[8] A. Batzios, C. Dimou, A. Symeonidis, P. Mitkas. BioCrawler: An intelligent crawler for the semantic Web. Expert Systems with Applications. v35, 2008, pp. 524–530.

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

[10] W. Dou, K. Lim, C. Su, N. Zhou, N. Cui. Brand Positioning Strategy Using Search Engine Marketing. MIS Quarterly. v34 n2, 2010, pp. 261-279.

[11] G. Boone, J. Secci, L. Gallant. Emerging Trends in Online Advertising. doxa comunicacion. v5 n5, 2009, pp. 241-253.

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

[13] R. Jaramh,T. Saleh, S. Khattab, I. Farag. Detecting Arabic spam Web pages using Content Analysis. International Journal of Reviews in Computing. v6, 2011, pp.1-8.

[14] M. Al-Kabi, H. Wahsheh, A. AlEroud et al. Combating Arabic Web spam Using Content Analysis. In Proceedings of the 2011 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT2011), Amman Jordan. (2011), pp. 401-404.

[15] H. Wahsheh, I. Alsmadi, M. Al-Kabi. Analyzing the Popular Words to Evaluate spam in Arabic Web Pages. IJJ: The Research Bulletin of JORDAN ACM – ISWSA. v2 n2, 2012a, pp. 22-26.

[16] M. Al-Kabi, H. Wahsheh, I Alsmadi, et al. Content Based Analysis to Detect Arabic Web spam. Journal of Information Science. v38 n3, 2012, pp. 284-296.

[17] H. A. Wahsheh, I. Abu Dosh, M. Al-Kabi, et al. Using Machine Learning Algorithms to Detect Content-based Arabic Web spam. Journal of Information Assurance and Security. v7 n1 ,2012b, pp.14-24.

[18] H. A. Wahsheh, M. N. Al-Kabi, I. M. Alsmadi.. Spam Detection Methods for Arabic Web Pages. First Taibah University International Conference on Computing and Information Technology (ICCIT 2012), Al-Madinah Al-Munawwarah, Saudi Arabia. v2, (2012c) pp 486-490.

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

[20] 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.

[21] W. Wang, G. Zeng. Content Trust Model for Detecting Web spam. In IFIP International Federation for Information Processing, (Etalle, S. and Marsh, S. Eds) Trust Management, 2007, pp. 139-152.

[22] A. Benczur, D. Siklosi, J. Szabo, et al. Web spam: a Survey with Vision for the Archivist. International Web Archiving Workshop (IWAW'08), Aaarhus, Denmark. (2008), pp. 1-9.

[23] J. Gadge, S. Sane, H. Kekre. Layered Approach to Improve Web Information Retrieval. Proceedings on 2nd National Conference on Information and Communication Technology NCICT. v7, (2011) pp. 28-32.

[24] T. Liu, J. Xu, T. Qin, et al. LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval. SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007), Amsterdam, Netherlands. (2007) pp. 1-10.

[25] C. Liang, L. Ru, X. Zhu. R-spamRank: A spam detection algorithm based on link analysis. Journal of Computational Information Systems. v3, 2007, pp. 1705-1712.

[26] Y. Chung, M. Toyoda, M. Kitsuregawa. Identifying spam link generators for monitoring emerging web spam. In Proceedings of the 4th

workshop on Information credibility (WICOW '10), Raleigh, North Carolina, USA. (2010), pp. 51-58.

[27] A. West, A. Agrawal, P. Bakeret et al. Autonomous link spam detection in purely collaborative environments. In Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11), ACM, Mountain View, California, USA. (2011), pp. 91-100.

[28] W. Zhang, D. Zhu, Y. Zhang, et al. Harmonic functions based semi-supervised learning for Web spam detection. In Proceedings of ACM Symposium on Applied Computing, Taichung, Taiwan. (2011), pp. 74-75.

[29] Y. Niu, Y. Wang, H. Chen, et al. A Quantitative Study of Forum spamming Using Context-based Analysis. In Proceedings of the Network & Distributed System Security (NDSS) Symposium, San Diego, California, USA. (2006), pp. 1-14.

[30] 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, Barcelona, Spain. v222, (2008), pp. 99-113.

[31] D. Saraswathi, A. Vijaya Kathiravan, S. Anita. A Novel Approach for Combating spamdexing in Web using UCINET and SVM Light Tool. International Journal of Innovative Technology and Creative Engineering. v1 n3, 2011, pp. 47- 52.

[32] E. Kumar S. Kohli. Improving Link spam Detection using spamizer, In Proceedings of the World Congress on Engineering and Computer Science 2011 (WCECS 2011), San Francisco, USA. v1, (2011) pp 19-21.

[33] M. Attia. Arabic Language Research and Translation. Retrieved March, 23, 2012 from the World Wide Web: http://attiaspace.com

[34] Y. Du, Y. Shi, X. Zhao. Using spam farm to boost PageRank. In the Proceedings of the 3rd international workshop on Adversarial information retrieval on the web AIRWeb '07, ACM. (2007), pp 29-36.

[35] J. Martinez-Romo, l. Araujo. Web spam Identification Through Language Model Analysis. Fifth International Workshop on Adversarial Information Retrieval on the Web AIRWeb ’09, Madrid, Spain. (2009) pp. 21-28.

[36] Y. Wang. A multinomial logistic regression modeling approach for anomaly intrusion detection. Computers & Security. v24, (2005) pp. 662-674.

[37] L. Yang. Distance Metric Learning: A Comprehensive Survey. Department of Computer Science and Engineering Michigan State University. 2006, pp. 1-51.

[38] D. Xhemali, C. Hinde,R. Stone. Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. International Journal of Computer Science. v4, 2009, pp. 6-23.

[39] H. Witten, E. Frank. Data Mining: Practica Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Systems, second edition, Morgan Kaufmann (MK). 2005, pp. 1-558.

Heider A. Wahsheh. Born in Jordan, in August 1987, he is a Master of Computer Information Systems at Yarmouk University in Jordan. He obtained his Master degree in Computer Information Systems (CIS) from Yarmouk University, Irbid-Jordan, 2012. His research interests include: Information Retrieval and Search Engines, Data Mining, Arabic Natural Language Processing, and Mobile Agent Systems.

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

