Journal of Information Assurance and Security.
ISSN 1554-1010 Volume 7 (2012) pp. 14-23
© MIR Labs, www.mirlabs.net/jias/index.html

Using Machine Learning Algorithms to Detect Content-based Arabic Web Spam

Heider Wahsheh[1], Iyad Abu Doush[1], Mohammed Al-Kabi[1], Izzat Alsmadi[1], and Emad Al-Shawakfa[1]

[1] Faculty of Information Technology and Computer Sciences, Yarmouk University,
P.O. Box 566, 21163 Irbid, Jordan
heiderwahsheh@yahoo.com
iyad.doush@yu.edu.jo
mohammedk@yu.edu.jo
ialsmadi@yu.edu.jo
Shawakfa@yu.edu.jo

***Abstract*: As the ranking of retrieved WebPages in Web search results is getting more important for several marketing purposes, many Web pages try to fool the search engines to get high ranks. This study aims to evaluate spam Web pages for pages with Arabic content using machine learning algorithms. Once spam techniques are applied, classifiers can be used to remove spam pages. The performed experiments are based on different training dataset sizes and extracted features. Two algorithms were then applied to detect spam pages, and compare between their different results. Results have showed that decision tree is better than Naïve Bayes in detecting Arabic spam pages.**

Keywords: Web spam, Web spam detection, Machine learning, Arabic content features.

I. Introduction and Background

While information is continuously expanding through the Internet, challenges for users to get what they are searching for are also continuously expanding. A spam email; either a document or information, are those emails that are received by users without their explicit acknowledgement. Such solicitation of spreading information or documents can have several purposes or reasons. The main reason is the marketing goal of trying to reach a large number of audiences. It may also be used to acquire or exploit user's information.

Web spam is not only about injecting or soliciting documents. It also includes incorrect ranking of pages in search engines' databases [1, 2, 3]. Many Websites and documents with low ranks are trying to deceive search engines using different methods in order to be ranked higher and be more visible to search engines [4]. For this, it is a major goal for search engines to counter Web spamming. To do this, search engines need to have several types of metrics that continuously evaluate Websites' relevancy and rank in order to be able to focus on Websites that get sudden high increase in relevancy or rank.

Search engines use several ranking metrics such as page rank (that depends on link popularity) and hits to evaluate Websites ranking. Spam detection methods use similar methods to detect unjustifiable increase of rank of some pages and documents [1, 2, 3, 5]. Web spam might be introduced in search engines' results via misleading users through the inclusion of spam URL's thus, increasing a Web page rank. For this, many algorithms for graph clustering were introduced to minimize such effects like that of [6, 7].

Web spammers continuously improve their hiding methods. Their goal is to avoid all detection techniques and to appear for spam detection tools as normal documents [5]. On the other hand, a major challenge for spam detection tools is to compromise between correctly detecting spam documents, while at the same time correctly detecting normal documents and keep them in the search results. In the next section, we will introduce the Receiver Operation Characteristics (ROC) or confusion matrix; a method used to evaluate prediction quality or accuracy.

Figure 1 shows the four attributes of the confusion matrix that combines the prediction outcome and the actual values.

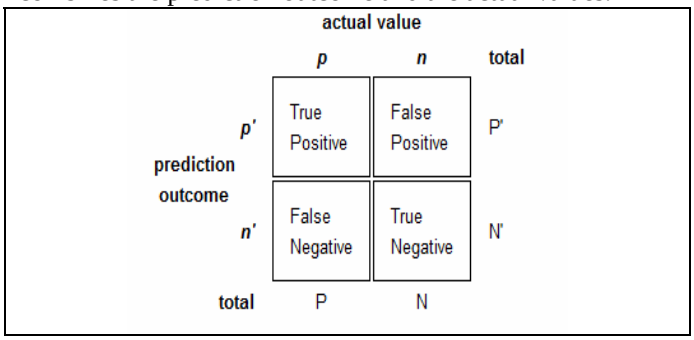

Figure 1. Confusion matrix

In spam detection problems, True Positive (TP) means that the document actual status is normal and it is identified as normal. True Negative (TN) means that the document is spam and is correctly identified as spam. For a spam detector to detect all documents correctly, TP and TN values should be 100%.

Their complements are false positive (FP), and false negative (FN) respectively [8].

Focusing only on one aspect and ignoring the rest may give wrong impressions about the spam detection abilities. For example, a spam detection tool may have a high TP value (e.g. 95%) which can be seen as a very reliable tool. However, this can be accompanied with a high FN value as well; which means that the tool is not good in detecting spams. On the other hand, a tool that detect all spams correctly (i.e. high TN), may also have high false alarms (i.e. high FP).
With the introduction of Web 2.0 platform and its applications, more of the Web spam has started to emerge through spambots. To fight such spam, researchers have started to work on tools and approaches for that purpose [9, 10, 11].

This paper studies the Arabic Web spam content-based detection, using two machine learning algorithms (decision tree, and Naïve Bayes). In our case study, we used a portion of the UK-2007 dataset, built an Extended-Arabic-2011 dataset, and recollected a new portion of the UK spam Web pages for evaluation purposes.

The rest of this paper is organized as follows: Section II presents a brief review of the related studies on Web spam detection, section III presents the classification methods used, section IV presents the conducted experiments; which include: data collections, features, proposed methodology, experiment results, and other comparisons. Last but not least section V presents the summary and concluding remarks.

II. Related Work

The process of content-based spam detection depends on several aspects related to the page content, and the search engine. Spammers struggle to keep their pages shown in high ranks of the search results. Several research papers have studied the problem of Web spam filtering, and have provided different techniques to deal with this problem.

In their study, Drost et al. [12] have exhibited how to identify a spam link, and discussed a method for generating training data. The conducted tests by the authors have revealed the effectiveness of classes of intrinsic and relational attributes, besides showing the effectiveness of contextual classifiers. In addition, the attributes of different Web pages were categorized according to their contribution to identify spam Web pages. Therefore, the authors have identified most discriminatory relational attributes.

A study by Kolari et al. [13] was dedicated toward the detection of Splogs; spam blogs which are used to deceive search engines by promoting the rank of their affiliated (i.e. target) Web sites to the spam blogs. In their study, Kolari et al. [13] have exhibited using a machine learning based approach, through SVM models, based on local and link-based features to identify the Splogs.

In another study Tian et al. [14] have presented how to detect Web spam pages using machine learning. Several human-engineered features were extracted from the raw data to be used by a semi-supervised learning to classify the unlabelled examples. Link-based features were also used in the training process. The test results have revealed the effectiveness of semi-supervised learning and the combinatorial feature-fusion method to improve the classifying ability of Web pages into either spam or non-spam.

Sydow et al. [5] have tried to exploit the Web pages linguistic features to discriminate between spam and non-spam Web pages. Different machine learning approaches were used to detect Web spam. Preliminary test results have shown the effectiveness of these linguistic features to identify spam Web pages. Sydow et al. [5] study has pointed to the importance of removing noisy data to purify the training dataset and therefore, have improved the classifier's accuracy. A study by Piskorski et al.[15] was based on the linguistic features, where over 200 linguistic based attributes were computed and studied, and the two well known Web spam corpora (i.e. Webspam-Uk2006 and Webspam-Uk2007) were used to evaluate the proposed attributes.

A novel algorithm; called WITCH, with a learning capability to identify spam hosts or pages was presented by Abernethy et al. [16, 17]. This algorithm had exploited the Web graph structure besides the contents and features of Web documents. The test results of WITCH have proved the effectiveness of their new algorithm in identifying spam Web documents. According to the authors, their algorithm had performed well, even with little training data. The authors have used the WEBSPAM-UK2006 spam collection as their dataset.

Geng et al. [18] have showed the capability of the machine learning based classifier to adapt to newly developed Spamdexing techniques. Therefore, they have proposed a two-stage classification strategy to detect spam Web documents. The proposed strategy was based on predicted spamicity of learning algorithms and hyperlink propagation. The preliminary tests have showed the effectiveness of their strategy. It was also noticed; according to authors, that more training data will lead to enhancing the effectiveness of the proposed strategy.

In order to improve the accuracy of spam classifiers, a research study by Dai et al. [19] has tried to benefit from content features within historical Web pages. Supervised learning techniques that are used in machine learning to produce a classifier, were used in their study to combine current page content classifiers with temporal features' classifiers. They have conducted tests using WEBSPAM-UK2007 dataset which have showed a 30% improvement relative to a baseline classifier; which only considers current page content.

Two Link based semi-supervised learning algorithms (i.e. Link-training and LS-training) were presented by Geng et al. [20]. The algorithms were based on the traditional self-training and topological dependency based link learning.

The algorithms have aimed at enhancing the effectiveness of the used classifiers to identify spams Web pages, where conducted tests have proved that these two algorithms were effective.

Hayati et al. [9] have conducted a study which was dedicated to the Spam problem within newly adopted Web 2.0 platforms, where Web spammers can host their materials in well-known Web sites such as social networking service sites and free encyclopedia. Using these new techniques; known as Web 2.0 Spam or Spam 2.0, have lead to a 50% increase in the amount of spam messages. Some of the spammers those days use Web spambots (Internet robots) to spread their spam content, where tests have showed that these spambots depend heavily on search engines to identify new target Web sites. Such epidemic techniques need exceptional solutions; for that, Hayati et al. [9] have presented a tool called (Honey Spam 2.0); based on tracking the behavior of Internet robots. Furthermore, researchers have performed a new study Hayati et al. [10], in which they embedded two action-based features sets (action time and action frequency) to help in identifying accurately spambots. This enhancement; and according to authors, has lead to an increased accuracy of 94.70% to identify the spambots.

The malicious Web acts have many facets. One of these is called Web spambots. A type of Web spider that spreads throughout the Web spam contents, and typically targeting Web 2.0 applications. The drawbacks of Web spambots are not restricted to the waste of the valuable resources of the Internet, but it has also mislead Internet users to unsolicited Websites; thus ranking spammers' campaign Websites higher than their actual ranks. Hayati et al. [11] have presented a novel way to utilize Web navigation behavior to detect Web spambots through an automated supervised machine learning solution. In addition, Hayati et al. [11] have proposed the usage of a new set of user behaviors; known as action set, to identify Web spambots. Web usage navigation behavior is used to create the feature set which is adapted to train Support Vector Machine (SVM) classifiers. The authors claim that their method of detecting Web spambots achieves a 96.24% accuracy.

The study of Niu et al. [21] has proposed the usage of genetic programming to detect Web spam. According to the authors, using genetic programming would lead to the establishment of systems with capabilities to adapt to the evolution of different Web spam techniques. Building Web spam classifiers that are capable to evolve and gain the best possible discriminating function, means as the spam techniques evolve by humans the classifiers would evolve automatically without any human intervention. The conducted tests on the proposed classifier have lead to a 26% improvement within the recall performance, an 11% improvement within the F-measure performance, and a 4% improvement within the accuracy performance, relative to SVM.

The study of Chung et al. [22] have exhibited an online learning algorithm that could be used to identify spam link generators, that can handle vast amount of data and many link-based features; including modified PageRank scores based on white and spam seeds, as well as neighboring host scores. Tests were conducted on a Japanese Web archive that was collected during three years, with 56% to 73% precision, with F-measure values of 0.54 to 0.68. Furthermore, the researchers have found that most of the new spam links were created by spam link generators.

A study by Metaxas [23] is different from the other presented studies in this section, since it tries to explore the influence of Web spam on the evolution of search engines, and it identifies strong relation between Web spamming methods and propagandistic techniques in society. Also, this researcher suggests an idea of propagating suspicion to a spamming network, if one of its Web pages is identified as a spam.

A novel semi-supervised learning algorithm; called HFSSL, was presented by Zhang et al. [24]. In this algorithm, the labeled and unlabeled Web pages were considered as vertices with a given weight for each Web page according to its similarity in a weighted graph. Tests on HFSSL have proved its effectiveness in detecting spammed Web pages.

A search engine results page (SERP) usually may present to their users a URL of Web spam pages, for this, Egele et al. [6] have conducted a study to detect URLs of spam Web pages within SERP. Egele et al. [6] have conducted comprehensive tests to discover the features that affect the ranking of any Web page within SERP, besides building a system based on the discovered features to remove spam links from SERP.

Largillier et al. [7] have studied the effects of node aggregation on Google's well known ranking algorithm PageRank, where a new graph clustering method was presented to reduce the effects of Web spamming. The authors of [7] have proved and have presented the necessary evidence to show the effectiveness of their method in detecting spam Web pages.

III. Classification methods

Different types of machine learning algorithms have been used for text classification [25, 26]. These algorithms were also used for spam filtering (e.g., [27, 28, 29]). Categorization of Web pages, as a spam and non-spam, is a supervised text classification problem. The classifier has to be trained with a group of Web pages that are categorized into either spam or non-spam pages [30].

As a proof of concept, we have used two classifier implementations of the machine learning toolkit Weka [31]: Decision tree, and Naïve Bayes. These classifiers were used for the classification of Web pages into either spam or non-spam pages.

Decision Tree is one of the common structures to organize the classification data; it visualizes what steps are taken to arrive at a classification decision. The decision is based on comparing values against some constants, through routing

from the start decision on the root node, until arriving at different paths based on different leaf nodes' attributes [32].

The J48 Decision Tree is one of the classification techniques available in WEKA. It represents information from a machine learning algorithm, offering high speed and powerful way to express the structure. It gives many options based on tree pruning; which can be used as a tool to correct potential problems over fitting, and used in many operational researches to identify the strategy needed to reach a specific goal [31]. With J48 Decision Tree, the data can be categorized as perfectly close as possible, which may ensure a maximum accuracy on the training data

Naïve Bayes belongs to a group of statistical techniques that use learning probabilistic knowledge; such as means and variances, that are also provided with different options in Weka. Examples of Weka NB options include: Naïve Bayes multinomial, simple, and Naïve Bayes updatable. It has been used to classify Web contents by applying the Bayes theorem with the assumption that all variables are conditionally independent [27, 28].

Naïve Bayes makes it easier to compute multiple variables. This makes it as one of the popular techniques for Web spam identification; providing a simple approach, clear semantics, very fast, and quite accurate results [31].

The formula of the Naive Bayes is [32]:

$$\Pr(S_n \setminus W) = \frac{\Pr(W \setminus S_n) \times \Pr(S_n)}{\Pr(W)} \quad (1)$$

Where

$\Pr(S_n \setminus W)$ = the prior probability of category *n*.

$\Pr(W \setminus S_n)$ = the conditional probability of the test page, given category *n*.

W = the new Web page to be classified.

IV. Experiments

The main goal of our experiments was to compare the efficiency of the machine learning algorithms using a case study of Web spam data collection. The proposed framework of the experiments includes three different data collections that estimate the classification accuracy.

A Data Collection

The lack of a benchmark collection of Arabic Web pages is still considered as one of the main problems affecting the research efforts in the field of Arabic Web spam filtering. In the literature, we only have found two of Arabic Web spam corpuses, mentioned in [33, 34]. We used and extended the dataset mentioned in [33].

The following three datasets of Web spam pages were considered and used in the experiments as well:

1) WEBSPAM-UK2007 dataset:

Castillo et al. [8] have offered the WEBSPAM-UK2007 dataset in their study. Their dataset was made available for public researchers, and was collected by laboratory volunteer at the University of Milan. The collection was manually labeled by human judges as to whether or not they are spam. We used a portion of the dataset in our experiment, around 4,000 Web documents; consisting of 2,000 Web spam sites and 2,000 non-spam Web sites. Figure 2 shows an example of UK spam Web site

2) UK-2011 Web spam dataset:

Due to the unavailability of some Web sites that were mentioned in WEBSPAM-UK2007, and the urgent need to compute new features, we have built a new dataset; called UK-2011, which was derived from the WEBSPAM-UK2007 dataset, to act as an alternative dataset.
Depending on the operational spam Web sites mentioned in WEBSPAM-UK2007 we have recollected the UK spam pages, by extracting spam pages from the available Web spam sites. The new dataset consisted of around 3,700 Web pages.

3) Extended-Arabic-2011 Web spam dataset

Wahsheh and Al-Kabi [33] have built a corpus of Arabic Web spam dataset, which has consists of 400 Arabic spam Web pages. This research has enhanced both the number of Arabic spam pages and their features. During this research we have expanded the collection to 10,000 Arabic spam pages, and have extracted some new features. During the time period from April 2011 to August 2011, we have manually labeled the pages as either spam or non-spam pages based on the authors' judgments and depending on different content-based features of the Web pages. Figure 3 shows an Arabic spam page example.

The used datasets in this study is available at https://sites.google.com/site/heiderawahsheh/home/Web-spam-2011-datasets .

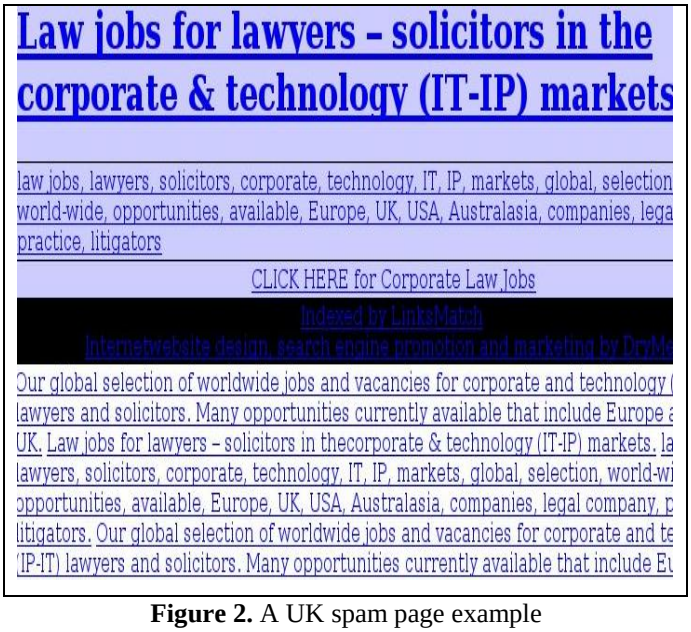

Figure 2. A UK spam page example

Figure 3. An Arabic Web spam example

B Features

Search Engine Optimization (SEO), is the process of identifying factors in the Web Pages which would impact search engine accessibility, including many elements of a Webpage that can affect the ranking algorithms thus leading to the highest possible visibility in search engine results. Spammers would always try to take advantage of the characteristics of SEO, to serve and improve their spam techniques [35].

The used features for the spam classification process are based on the content of Arabic or English Web pages, which involve the features used in [33], as well as the common features that were proposed by [36, 37, 38, and 39]. In addition, we have proposed the following new features to be used according to the characteristics of Arabic Web spam detection as following:

1) Total number of words in the <Meta> tag (Meta description): Key stuffing is a practice in which keywords within HTML elements are repeated too many times. Spammers use the stuffing practice in the Meta tag, which aims to embed the content of Arabic Web pages with a number of popular words. These words are irrelevant to the meaning of the content of the Arabic Web page.

2) Minimum and maximum words' length in Web pages. Spammers try to increase the length of the popular and important keywords in the Web page to increase its rank. In order to identify this feature we need to know both the minimum and the average words length in the non-spam Web pages.

3) The Total number of images in the Web page. Increasing the number of images in a Web page can leads to attracting more users. This could increase the rank of a spam Web page in the search results.

Table 1, at the end of the paper, shows all features used in this research along with the SEO guidelines for each feature. The fields with "Cannot determine" value indicates that this property cannot be used as a standalone feature to identify a Web page as either spam or non-spam. For example, the number of words in the Web page is connected with the property of the number of different words in the Web page.

C Proposed Methodology

In this study, we have used two machine learning algorithms (decision tree and Naïve Bayes) to combat content-based Arabic Web spam. We have built an Arabic Web spam dataset that includes 10,000 Arabic spam Web pages. In addition, we have used an updated version of the WEBSPAM-UK2007; called UK-2011, dataset with 3,700 spam pages.

To achieve the research goals, we have divided the work into four steps:

1) Compute Common Features for Web Spam Detection Independent Language. Based on the classification features, which were proposed in [36, 37, 38, and 39], and other proposed features presented in Table 1, we have used the common features that were already computed in WEBSPAM-UK2007 dataset. We have also computed common features for both Extended-Arabic-2011 Web spam dataset and the UK-2011. After that, we have applied the two machine learning algorithms (i.e., decision tree and Naïve Bayes) on the three datasets. Finally we compared the obtained results of detecting Web spam.

2) We have used the Arabic Spam Detection Features (ASDF); proposed in [33], to compare between the Arabic-2011 and UK-2011 datasets.

3) We have also computed the proposed new features on the Extended-Arabic-2011 Web spam dataset, and the new UK-2011 dataset, and then compared the results.

4) Finally, all features were merged as one group of Web Spam Detection features and then compared the Arabic Web spam 2011 dataset and the UK-2011 dataset. This approach helps us to reach the most appropriate features needed to detect Web spam pages.

D Experimental Results

Two of the machine learning algorithms were applied using one Arabic and one English datasets. We further have computed the error and accuracy percentages for each of the algorithms at each step. For this, we have computed the Kappa Statistic (KS), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). We also have used the ROC metrics: True Positive (TP), False Positive (FP), Precision (P), Recall (R), and F-Measure (F-M). The following shows the two used machine learning algorithms in this research:

1) Decision Tree algorithm

We have applied the J48 Decision Tree algorithm. First we have computed the common features with WEBSPAM-UK2007; the tree consists of 35 leaves, and 69 nodes. The correctly classified instances were 3223, constituting 75.1632% accuracy, with 24.8368% error rate indicating 1065 incorrectly classified instances.

We have also computed the common features using the Arabic-2011 Web spam dataset, and have applied the J48 Decision Tree algorithm, which consisted of 43 leaves and 85 nodes. The correctly classified instances were 3382, with total of 99.4706% accuracy, 0.5294% error rate with only 18 incorrectly classified instances.

We also have computed the common features using the UK-2011 dataset and have applied the J48 Decision Tree. The resulted tree consisted of 242 leaves and 483 nodes. Experiments have achieved 79.846% accuracy and 20.154% error rate, with 3007 correctly classified instances, and 759 incorrectly classified instances.

From the above results, we can notice that the new UK-2011 dataset has provided a closer accuracy percentage of that of the WEBSPAM-UK2007. This shows the reliability of the new UK-2011 dataset; especially, for computing some new features.

We have applied the second step of our methodology; compute (ASDF), using both Extended-Arabic-2011 Web spam dataset and UK-2011 dataset. With the new UK-2011 dataset, we have achieved 82.0234% accuracy with 3089 correctly classified instances and 17.9766% error rate with 677 incorrectly classified instances. As for the Extended-Arabic-2011 Web spam dataset however, we have achieved 98.1471% accuracy with 3337 correctly classified instances, and 1.8529% error with 63 incorrectly classified instances.

Furthermore, we have applied the third step of our methodology, to compute the proposed new features. Using the UK-2011 dataset, we have achieved 85.316% accuracy and 14.684% error, with 3213 correctly classified instances, and 553 incorrectly classified instances.

As for the Extended-Arabic-2011 dataset, we have achieved 98.7101% accuracy and 1.2899% error, with 9872 correctly classified instances, and 129 incorrectly classified instances.

Finally, we have applied the fourth step of our methodology, to compute Web spam detection features on the UK-2011 dataset. The tree has consisted of 216 leaves and 431 nodes. The method has achieved 88.1306% accuracy with 3319 correctly classified instances and 11.8694% error rate with 447 incorrectly classified instances.

With the Extended-Arabic-2011 dataset, the tree has 57 leaves and 113 nodes. Applying the step on the dataset, the method has achieved 99.0882% accuracy with 3369 correctly classified instances, and 0.9118% error with 31 incorrectly classified instances.

Tables 2 and 3 show the ROC metrics for the UK-2011 and the Extended-Arabic-2011 dataset, while Table 4 shows the dataset statistical information.

Table2. ROC metrics of UK-2011.

Class	**TP**	**FP**	**P**	**R**	**F-M**	**ROC**
Spam	0.89	0.132	0.884	0.893	0.889	0.89
Non-spam	0.86	0.107	0.874	0.868	0.873	0.89
Weight-ed AVG	0.88	0.12	0.88	0.88	0.88	0.89

Table 3. ROC metrics of Arabic-2011 dataset.

Class	**TP**	**FP**	**P**	**R**	**F-M**	**ROC**
Spam	0.99	0.01	0.989	0.992	0.99	0.99
Non-spam	0.98	0.008	0.992	0.989	0.991	0.99
Weight-ed AVG	0.99	0.009	0.991	0.99	0.991	0.99

Table 4. Dataset statistic information.

Dataset	**KS**	**MAE**	**RMSE**	**RAE**	**RRSE**
UK 2011	0.761	0.135	0.329	27.2%	66.0%
ARABIC – 2011	0.981	0.010	0.0947	2.09%	18.49%

2) Naïve Bayes algorithm

We have also applied the methodology steps using the Naïve Bayes algorithm. First we have computed the common features using the WEBSPAM-UK2007 dataset. The correctly classified instances were 2,513 and 58.6054% accuracy, with 41.3946% error representing 1,775 incorrectly classified instances.

Then we have computed the common features using the Extended-Arabic-2011 Web spam dataset, the correctly classified instances were 2,753, 80.9706% accuracy, and 19.0294% error with 647 incorrectly classified instances.

In addition, we have computed common features using the UK-2011 dataset. The method has achieved 55.7886% accuracy and 44.2114% error, with 2,101 correctly classified instances, and 1,665 incorrectly classified instances.

Applying the Naïve Bayes algorithm using the UK-2011 dataset has showed that the new UK-2011 dataset has provided a closer accuracy percentage with WEBSPAM-UK2007 dataset. These results confirm the results yielded from applying the decision tree algorithm on the same datasets.

Applying the second step of our methodology to compute (ASDF), we use both Extended-Arabic-2011 Web spam and the UK-2011 datasets, which achieved similar results to those

yielded by decision tree algorithm. With the new UK-2011 dataset however, we have achieved 54.0361% accuracy with 2,035 correctly classified instances and 45.9639% error with 1,731 incorrectly classified instances.
On the other hand, using the Extended-Arabic-2011 Web spam dataset, we have achieved 82.3235% accuracy with 2,799 correctly classified instances, and 17.6765% error with 601 incorrectly classified instances.

We have also applied the third step of our methodology and computed the proposed new features using the UK-2011 dataset. The method has achieved 55.1779% accuracy and 44.8221% error, with 2,078 correctly classified instances, and 1,688 incorrectly classified instances.

With the Extended-Arabic-2011 dataset however, the method has achieved 75.6724% accuracy and 24.3276% error, with 7,568 correctly classified instances, and 2,433 incorrectly classified instances.

Finally, applying the fourth step of our methodology, which compute Web spam detection features on the UK-2011 dataset. This method has achieved 56.8508% accuracy with 2,141 correctly classified instances and 43.1492% error with 1,625 incorrectly classified instance.

With the Extended-Arabic-2011 dataset the method has achieved 83.6176% accuracy with 2,843 correctly classified instances, and 16.3824% error with 557 incorrectly classified instances.

Tables 5 and 6 show the ROC metrics for the new UK-2011 and the Extended-Arabic-2011 datasets. On the other hand, Table 7 presents the dataset statistical information.

Table 5. ROC metrics of UK-2011.

Class	**TP**	**FP**	**P**	**R**	**F-M**	**ROC**
Spam	0.97	0.895	0.553	0.979	0.707	0.52
Non-spam	0.10	0.021	0.815	0.105	0.185	0.52
Weighted AVG	0.56	0.485	0.676	0.569	0.462	0.52

Table 6. ROC metrics of Arabic-2011 dataset.

Class	**TP**	**FP**	**P**	**R**	**F-M**	**ROC**
Spam	0.95	0.285	0.771	0.958	0.854	0.94
Non-spam	0.71	0.042	0.944	0.715	0.813	0.94
Weighted AVG	0.83	0.164	0.857	0.863	0.834	0.94

Table 7. Dataset statistical information.

Dataset	**KS**	**MAE**	**RMSE**	**RAE**	**RRSE**
UK 2011	0.088	0.434	0.652	87.1%	130.74%
ARABIC - 2011	0.672	0.175	0.371	35.0%	74.25%

The obtained results show that using the Extended-Arabic-2011 dataset to detect Web spam in Arabic Web pages gave better results than the UK-2011 dataset. Furthermore, the results from the experiments have proved that the decision tree algorithm is better than Naïve Bayes in terms of accuracy and the ability to correctly detect an Arabic Web spam page.

We have compared the results of this paper with those of [33] and [34] which studied Arabic Web spam. The results of this paper have showed that the Decision Tree algorithm is a good technique to be used in order to identify an Arabic Web spam. This confirms what has been yeilded by Jaramh et al. [34] study which shows that the decision tree algorithm is the best algorithm to be used to identify Arabic Web spam pages.

V. Conclusions

Web spam is any manipulation of Web pages that aims to mislead the ranking algorithms of search engines.

Arabic Web spam has become a very serious problem to the Arab Internet community. It is important to be able to filter Web pages into either spam or non-spam to rank them correctly in the search results.

In this paper, we have collected around 14,000 spam Web pages that were represented by two datasets with Arabic and English language content. We have presented the usage of two machine learning algorithms: Naïve Bayes and Decision Tree. Then we have applied these algorithms on the datasets. The experimental results have showed that the decision tree classifier is more sensitive to the detection of Arabic Web spam pages relative to other classification algorithms.

The obtained results show that there is a better classification accuracy using Arabic dataset. More accurate results in detecting Web spam pages have been obtained using the proposed Extended-Arabic-2011 dataset compared with the UK-2011 dataset.

References

[1] Z. Gy¨ongyi, and H. Garcia-Molina. “Spam: It’s not just for inboxes anymore”, *IEEE Computer Magazine,* 38 (10), pp. 28–34, 2005.

[2] M. R. Henzinger, R. Motwani, and C. Silverstein. “Challenges in Web search engines”. *SIGIR Forum*, 36 (2), pp. 11–22, 2002.

[3] A. Singhal. “Challenges in running a commercial search engine”, *In IBM Search and Collaboration Seminar.* IBM Haifa Labs, pp. 432, 2004.

[4] Z. Gyongyi and H.Garcia-Molina. “Web spam taxonomy”. *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web*, Chiba, Japan, pp. 1-9, 2005.

[5] M. Sydow, J. Piskorski, D. Weiss, and C. Castillo. "Fighting Web Spam", *in Mining Massive Datasets for Security*, D.Perrotta et al. (eds), NATO Science for Peace and Security Series, D: Information and Communication Security, 19, pp. 134-153, Amsterdam, Netherlands, 2008.

[6] M. Egele, C. Kolbitsch and C. Platzer. "Removing Web spam links from search engine results", *Journal in Computer Virology*, 7 (1), pp. 51–62, 2011.

[7] T. Largillier, and S. Peyronnet. "Webspam demotion: Low complexity node aggregation methods", *Neurocomputing Journal*, 76 (1), pp. 105–113, 2012.

[8] C. Castillo, D. Donato, A. Gionis, V. Murdock, F. Silvestri. "Know your neighbors: Web spam detection using the Web topology". *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 423-430, 2007.

[9] P. Hayati, K. Chai, V. Potdar, and A. Talevski. "HoneySpam 2.0: Profiling Web Spambot Behaviour". *Lecture Notes in Computer Science (LNCS), Principles of Practice in Multi-Agent Systems (PRIMA)*, Nagoya, Japan, pp. 335–344, 2009.

[10] P. Hayati, K. Chai, V. Potdar, and A. Talevski, "Behaviour-Based Web Spambot Detection by Utilising Action Time and Action Frequency". *In Proceedings of ICCSA (2)*, pp.351-360, 2010.

[11] P. Hayati, V. Potdar, K. Chai, and A. Talevski. "Web Spambot Detection Based on Web Navigation Behaviour". *In: Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications*, pp. 797-803, 2010.

[12] I. Drost, and T. Scheffer. "Thwarting the Nigritude Ultramarine: Learning to Identify Link Spam". *In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 233-243, 2005.

[13] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. "Detecting spam blogs: A machine learning approach". *In: Proceedings of the National Conference on Artificial Intelligence* (*AAAI*), Boston, MA, USA, 2, pp. 1-6, July 2006.

[14] Y. Tian, G. M. Weiss, Q. Ma. "A Semi-Supervised Approach for Web Spam Detection using Combinatorial Feature-Fusion". *In: Proceedings of the ECML/PKDD 2007 Graph Labeling Workshop and Web Spam Challenge*, pp. 16-23, 2007.

[15] J. Piskorski, M. Sydow, and D. Weiss. "Exploring linguistic features for Web spam detection: a preliminary study". *In: Proceedings of the 4th international workshop on Adversarial information retrieval on the Web* (*AIRWeb '08*), pp. 25-28, 2008.

[16] J. Abernethy, O. Chapelle, C. Castillo. "Web spam Identification Through Content and Hyperlinks". *AIRWeb '08*, pp. 1-4, 2008.

[17] J. Abernethy, O. Chapelle, and C. Castillo. "Graph regularization methods for Web spam detection", *Machine Learning*, 81 (2), pp. 207–225, 2010.

[18] G. Geng, C. Wang, and Q. Li. "Improving Spamdexing Detection Via a Two-Stage Classification Strategy". *In: Proceedings of the 4th Asia Information Retrieval Symposium* (*AIRS 2008*), pp. 356-364, 2008.

[19] N. Dai, B.D. Davison and X. Qi. "Looking into the Past to Better Classify Web Spam". *In: Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web* (*AIRWeb'09*), pp. 1-8, 2009.

[20] G. G. Geng, Q. Li, and X. Zhang. "Link based small sample learning for Web spam detection". *In: World Wide Web Conference Series*, pp. 1185-1186, 2009.

[21] X. Niu, J. Ma, Q. He, S. Wang and D. Zhang. "Learning to Detect Web Spam by Genetic Programming". *In: Proceedings of the 11th international conference on Web-age information management* (*WAIM'10*), pp. 18–27, 2010.

[22] Y. J. Chung, M. Toyoda, and M. Kitsuregawa. "Identifying Spam Link Generators for Monitoring Emerging Web Spam". *In: World Wide Web Conference Series*, pp. 51-58, 2010.

[23] P. T. Metaxas. "Web Spam, Social Propaganda and the Evolution of Search Engine Rankings", *Web Information Systems and Technologies*, 45, pp. 170–182, 2010.

[24] W. Zhang, D. Zhu, Y. Zhang, G. Zhou, and B. Xu. "Harmonic functions based semi-supervised learning for Web spam detection", *In: ACM Symposium on Applied Computing*, pp. 74-75, 2011.

[25] J. Su and H. Zhang. "Full Baysian Network Classifiers". *In the Proceedings of 23 rd International Conference on Machine Learning*, Pittsburgh, PA, pp. 1-8, 2006.

[26] B.Yu and Z.Xu. "A Comparative Study for Content-Based Dynamic Spam Classification Using Four Machine Learning Algorithms", *Knowledge-Based Systems*, 4 (21), pp. 355-362, 2008.

[27] H. Zhang and D. Li. "Naïve Bayes Text Classifier". *In the Proceedings of the IEEE International Conference on Granular Computing*, pp. 708-711, 2007.

[28] G. paul. "Better Bayesian Filtering". *In the Proceedings of the 2003 spam conference*, 2 (3), pp. 24-30, Jan 2003.

[29] C. Castillo, D. Donato, L. Becchetti1, P. Boldi, S. Leonardi, M. Santini and S. Vigna. "A reference Collection for Web Spam". *SIGIR Forum*, 40, pp. 11-24, 2006.

[30] H. Najadat and I. Hmeidi. "Web Spam Detection Using Machine Learning in Specific Domain Features", *Journal of Information Assurance and Security*, 3, pp. 220-229, 2008.

[31] I. H. 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.

[32] D. Xhemali, C. J Hinde, and R. G Stone. "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages", *International Journal of Computer Science*, 4, pp. 16-23, 2009.

[33] H. A. Wahsheh, M. N. Al-Kabi. "Detecting Arabic Web Spam". *The 5th* International *Conference on*

Information Technology, ICIT 2011, Paper ID (631), pp. 1-8, 2011.

[34] R. Jaramh, T. Saleh, S. Khattab, I. Farag. "Detecting Arabic Spam Web pages using Content Analysis", *International Journal of Reviews in Computing*, 6, pp. 1-8, 2011.

[35] J. Zhang, and A. Dimitroff. "The impact of Webpage content characteristics on Webpage visibility in search engine results (Part I)", *Information Processing and Management*, 41, pp. 665-690, 2005.

[36] W. Wang, G. Zeng, M. Sun, H. Gu, Q. Zhang. "EviRank: An Evidence Based Content Trust Model for Web Spam Detection". *APWeb/WAIM*, pp. 299-307, 2007.

[37] W. Wang, G. Zeng, D. Tang. "Using evidence based content trust model for spam detection", *Expert Systems with Applications*, 37 (8), pp. 1-8, 2010.

[38] W. Wang, G. Zeng. "Content Trust Model for Detecting Web Spam". *IFIP International Federation for Information Processing.* pp. 139-152, 2007.

[39] A. Ntoulas, M. Najork, M. Manasse, D. Fetterly. "Detecting Spam Web Pages through Content Analysis". *Proceedings of the World Wide Web Conference*. pp. 83-92, 2006.

[40] J. Zhang, and A. Dimitroff. "The impact of metadata implementation on Webpage visibility in search engine results (Part II)", *Information Processing and Managemen*t, 41, pp. 691-715, 2005.

Author Biographies

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 bachelor degree in Computer Information Systems (CIS) from Yarmouk University, Irbid-Jordan, 2005. His research interests include: Information Retrieval and Search Engines, Data Mining, and Mobile Agent Systems.

Iyad Abu Doush is an Assistant Professor in the Department of Computer Science at Yarmouk University, Jordan. Dr. Abu Doush has born in Amman/Jordan in 1979. He received his BSc. In Computer Science from Yarmouk University, Jordan, 2001, and his M.S. from Yarmuk University, Jordan, 2003. He earned his Ph.D. from the Computer Science Department at New Mexico State University, USA, 2009. Since then, he has been a professor of computer science, Yarmouk University, Jordan. His research interests include assistive technology, intelligent interfaces, and multi-modal interfaces. Other research interests include human computer interaction, computational intelligence, and collaborative virtual environments.

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 Professor in the Department of Computer Information Systems at Yarmouk University. Prior to joining Yarmouk University, he spent six years at 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 , USA in the years 2006 and 2008 respectively. His main areas of research include: software engineering, testing, metrics, and information retrieval.

Emad Al-Shawakfa was born in Irbid/Jordan in 1964 is an Assistant Professor at the Computer Information Systems Department, Faculty of IT, Yarmouk University since September 2000. Dr. Al-Shawakfa holds a PhD degree in Computer Science from Illinois Institute of Technology (IIT) – Chicago, USA in the year 2000, a M.Sc. in Computer Engineering from Middle East Technical University – Ankara, Turkey in the year 1989, and a B.Sc. in Computer Science from Yarmouk University – Jordan in the year 1986. Dr. Al-Shawakfa research interests are in Computer Networks, Data Mining, and Natural Language Processing. His PhD research was about creating natural language interfaces to Network Operating Systems using the concept of case frames; the first to apply the concept to the Arabic natural language processing. His research was based on work experience in networking at IIT. He has many publications in the research fields and currently working on others.

Table 1. Features of Web spam detection.

Feature Name	SEO Guidelines	Status of Spam/Non-spam Web pages
1. The number of Arabic-English words in the title of Web pages.	Duplicated keywords in the title increase the visibility in search engine results. The threshold used is up to three duplications, if it exceeds three there is a downturn in terms of visibility [35].	In Extended-Arabic dataset, the average number of Arabic-English words in the title of Web pages in spam Web pages around 10.38, while the average in the non spam Web pages was around 5.96.
2. The number of the (Arabic and/or English) words in the Web pages.	Duplicated keywords in the full-text of a Web pages, increases the visibility in search engine results (no limited boundary of this feature) [35].	Cannot determine.
3. The average of Arabic-English words lengths in the Web pages.	Prolong the size of the word, to introduce composite words; based on concatenate 2-4 words into one word (e.g., free mp3 video). It can be used with English and Arabic language [37, 39].	Cannot determine.
4. The number of different words in the Web pages.	By repeating specific keywords in the Web page the spammers hope to raise the rank of the Web page in the search results [35].	Cannot determine.
5. The amount of anchor text in the Web pages.	Higher fractions of anchor text may imply higher prevalence of spam [37].	In Extended-Arabic dataset, the average in spam Web pages around 110.84, while the average in the non spam Web pages around 94.72.
6. The number of Arabic-English words in meta tag.	Keywords in metadata should come directly from the Webpage. Web pages with metadata elements achieve better visibility performance than those without metadata elements [40].	In Extended-Arabic dataset, the average in spam Web pages around 60.33, while the average in the non spam Web pages around 35.68.
7. The minimum Arabic-English word length inside the Web pages.	We assume that the Minimum Arabic–English words consist of 3 characters [39] in the conducted experiments.	Cannot determine.
8. The maximum Arabic-English word length inside Web pages.	Increase the size of the word, to introduce composite words. This is performed by concatenating 2-4 words into one word (e.g., freemp3video). It can be used with English and Arabic language [37, 39].	Cannot determine.
9. The number of images in the Web pages.	Increase the number of images to attract more users. This could increase the rank of the spam Web page in the search results.	In Extended-Arabic dataset, the average in spam Web pages around 70.24, while the average in the non spam Web pages around 37.46.
10. The number of characters in the Meta tag.	Words and characters in the metadata elements extracted from title and full-text, achieve better visibility performance than only characters or keywords extracted only from full-text [40]. The spam Web page increase the number of words and characters used in the metadata to increase the rank of the Web page in the search results.	In Extended-Arabic dataset, the average in spam Web pages around 452.32, while the average in the non spam Web pages around 438.35.
11. The Compression rate of the Web pages.	Increasing the compression rate in a Web page to hide redundant content [39].	In Extended-Arabic dataset, the average in spam Web pages around 63%, while the average in the non spam Web pages around 35%.