DETECTING ARABIC SPAM WEB PAGES USING CONTENT ANALYSIS

1RADWAN JARAMH, 2TALAL SALEH, 1SHERIF KHATTAB, 1IBRAHIM FARAG
1 Computer Science Dept., Faculty of Computers and Information, Cairo University, Egypt
2 Computer Science Dept., Faculty of Computer Science and Engineering, Hodeida University, Yemen

ABSTRACT

In this paper, we propose a set of new features to enhance the classification of Arabic Web pages into spam and non-spam under different classification algorithms, namely Decision Tree, Naïve Bayes, and LogitBoost. We compare our features, which we call Arabic Content Analysis (ACA) features, to state-of-the-art Content Analysis (CA) features for spam detection in the English Web. We show that augmenting the CA features with our ACA features achieves an increase in detection accuracy of Arabic spam pages compared to CA features alone. When combined, ACA and CA features correctly identified 5,536 pages of the 5,645 Arabic spam pages that we used for testing with a false positive rate of 1.9% using the Decision Tree classifier. We also identified the top-ranked features using the Gain Ratio method.

Keywords: Web Spam, Web Pages, Arabic Web Spam, Detecting Arabic Spam, Arabic Corpus, Arabic Keywords, Spamdexing.

1. INTRODUCTION

Due to increased trading through the Web, the U.S online retail transactions have grown to reach $175 billion in 2007, and it is projected to grow to $335 billion by 2012 [16]. However, spam pages are used to manipulate search engines to give false high ranking to certain pages in popular searches, exploiting the fact that approximately 85% of the search queries result in a click on a link in the first page of results [17].

As such, spam pages lead to user frustration, distortion of the results, information pollution, in general, and slow searching. Fig. 1 shows an example of an Arabic spam page that contains stuffing keywords (e.g. البنك السعودي الفرنسي (French Saudi Bank), السعودي (Saudi), الفرنسي (French), which are useless to a human viewer.

In this paper, we propose a set of new features that enhance the classification of spam Arabic Web pages. We call our features Arabic Content Analysis (ACA) features.

Our ACA features include cosine similarity among components of a Web page (e.g. document body, title, keywords, and description), fraction of page words drawn from the top globally popular words, count of irregularly long words, independent and conditional n-gram likelihoods [3], and number of excessively repeated words.

To test our features, we collected a corpus of Arabic Web pages containing 12747 pages, 5645 spam and 6602 non-spam. We collected them from February 2010 to July 2010. The corpus was
collected manually using search engines (e.g. Google, Bing, AltaVista, Maktub, and Ayna). We relied on international and Arabic search engines. We also used Arabic directory sites, statistics on Arabic keywords, Arabic spam sites, and forums.

We experimentally compared three groups of features: our ACA features, Content Analysis (CA) features proposed in [3] for spam detection in the English Web, and ACA combined with CA (ALL) features. We used three classifier techniques, namely Decision Tree, Naïve Bayes, and LogitBoost. We also used the gain ratio technique to identify the most significant features. The highest F-measure [29] value was 0.98 with a false positive rate of 0.014 for ALL features using the Decision Tree classifier.

This paper is structured in six sections. In Section 2, we describe background on Web spam detection, and discuss related work. In Section 3, we present our Arabic Web corpus, which contains spam and non-spam pages. In Section 4, we describe our ACA features and divide them into four groups, namely globally popular words, counting, cosine similarity, and n-gram syntax features. In Section 5, we describe our experimental results for ALL features, ACA features, and CA features with three classifiers, and we discuss the results of our experiments. Finally, in Section 6, we present conclusions and future work.

2. BACKGROUND AND RELATED WORK

In this section we present some background on the problem of Web spam detection and discuss related work.

2.1 Web Spam Detection

To deceive search engines, spammers do many tricks; some of them depend on modifying page contents. In content-based Web spam, a popular method is to repeat popular words in the page, and make them invisible to the user but visible to search engines. Example popular words are facebook (facebook), google (غوجل), yahoo (ياهو), travian (تري반), hotmail (هتاميل), photo (صور), youtube (يوتوب), news (الأخبار), games (الألعاب), and search (البحث). These popular words fill the body and the title of a spam Web page. We note that each English popular word has four forms: written in English, written in Arabic, written in English while the language of the operating system is Arabic, and written in Arabic while the language is English.

Whereas spam pages are usually useless for the users, this spamming technique deceives the search engines and makes the page occupy a high position in the top search results of popular searches.

2.2 Related Work

In this subsection, we describe previous studies on Web spam detection and classifiers.

Content-based detection. Ntoulas et al. [3] investigated content-based Web spam detection. They used a collection of 2,364 Web spam pages. Their features include average length of words, amount of anchor text, fraction of page drawn from globally popular words, and independent n-gram likelihoods. They used the Decision Tree classifier. Their method achieved an F-measure of 0.862 and a false positive rate of 1.3%.

H. Wahsheh et al. [23] applied three classifiers; namely Decision Tree, Naïve Bayes, and nearest neighbors; and their features included the number of words in page, average sentence length in words, and complexity factor of Web page within lexical density. They used a collection of 402 Arabic Web pages: 202 spam pages and 200 non-spam. They achieved an accuracy of 96.8% and a false positive rate of 3.125%. Fetterly et al. [5] presented techniques to identify duplication (by copy-and-paste) and tested their techniques on a collection of 151 million Web pages collected in December 2002.


Link-based detection. Yates et al. [7] presented a spam technique to raise google's PageRank of a page. Gyongyi et al. [34] studies how to detect link farms (i.e., sites exchanging links for mutual benefit). Benečíř [18] proposed a technique based on cosine similarity to increase the effectiveness of their classifier.

Hybrid (content-link) detection. Gyongyi et al. [1] studied a variety of commonly used Web spamming techniques and proposed some heuristic
spam techniques, such as Boosting (hiding to manipulate search engine results). The studied hiding techniques include content hiding, cloaking, and redirection.

Henzinger et al. [2] studied the impact of Web spam, such as link spam, doorway pages, and cloaking. Webb et al. [12] analysed content and HTTP headers on the Web Spam Corpus of about 350,000 Web spam pages. Liu [11] analysed content hiding by using the same color of background for the font and hyperlinks.

3. DATA COLLECTION

We built a corpus of Arabic Web pages containing 12747 pages, 5645 spam and 6602 non-spam. We collected them from February 2010 to July 2010. The corpus was collected manually using search engines (e.g., Google, Bing, AltaVista, Maktoob, and Ayna). We relied on international and Arabic search engines, because the latter are more specialized in searching for Arabic keywords and use less spam-tolerant page-ranking algorithms, which helped us find a larger number of spam pages.

We relied on Arabic keywords such as بيوتوب، تويتر، الجزيرة، عربي، إعلام، عربي، العربي، البرامج، عرب. We also used Arabic directory sites, statistics on Arabic keywords, known Arabic spam sites, and forums. Examples of each spam category are listed in Table 1. We collected and analyzed the html source code of each page.

Table 1. Some domains that contained Arabic spam Web pages.

http://www.arab2m.com/vb/showthread.php?p=45604
http://www.vb-4.com/vb

For non-spam pages, we collected them from several trusted sites, which include for instance government, education, newspapers, and other cultural sites. All of these sites are legitimate and trusted due to their good reputation and official content. Such official Web sites are usually well established and classified by Google and Alexa as trusted with high ranks and therefore have a high page rank, up to eight on Alexa (e.g., Aljazeera [24], Islam online [25], Al-Ahram newspaper [26], Attorney General [27], and Supreme Judicial Council Yemen [28]).

4. OUR ARABIC CONTENT ANALYSIS (ACA) FEATURES

In this paper, we propose a set of features, which we call Arabic Content Analysis or ACA in short, to enhance the accuracy of detecting Arabic spam Web pages. We combine ACA with CA features [3] and call the combination ALL features. We divide the ALL features into four groups, namely globally popular words, counting, cosine similarity, and syntax n-gram. Each group contains a number of features for detecting Arabic spam Web pages.

We conducted experiments to measure the accuracy of each group and each feature. We used different classifier techniques, namely Decision Tree, Naïve Bayes, and LogitBoost.

In the following subsections, we start by listing all the features then we describe each feature group.

4.1 Overview

Our ACA features include cosine similarity of the following: document with title, document with keywords, document with description, keywords with document, keywords with title, keywords with description, document with keywords, description with document, description with keywords, description with title.
The ACA features also include the fraction of page words drawn from one hundred globally popular words, fraction of page words drawn from fifty globally popular words, fraction of globally popular words in the page, count of words with length > 15 characters, independent 2-gram likelihoods [3], conditional 2-gram likelihoods [3], and number of words repeated in page >= 10 times.

4.2 Globally Popular Words

Globally popular words are those extensively searched for by search engines. A lot of Web pages exploit these globally popular words to mislead search engine page-ranking algorithms to obtain high ranks in the first page of search results. Often, these words are used in spam Web pages to obtain high ranks in the first page of search results. We note that some of these keywords contain meaningless sequences of English characters and symbols. This is a result of typing Arabic words using an English keyboard or vice versa.

We have collected 100 globally popular words based on search statistics [19]. Examples of these words are: (youtube), (travian), (facebook), (twitter), (game), (_proxy), (google), (song), (news), (Arab), (hotmail), (video), (music), (.proxy), (facebook), ( inflammation), (hotmail), (news), (facebook), (google), (news), (yahoo), (youtube), (hotmail), (facebook)

We note that some of these keywords contain meaningless sequences of English characters and symbols. This is a result of typing Arabic words using an English keyboard or vice versa.

This feature group includes two types of heuristics: fraction of page words drawn from the globally popular words and fraction of the globally popular words that exist in the page. The first heuristic is defined as the fraction of globally popular Arabic words from all words in the page, or more precisely, number of page words that are globally popular / total number of words in page. The second heuristic is defined as the number of globally popular Arabic words used in a page divided by one hundred (number of globally popular words), or, number of globally popular words contained in a page / total number of globally popular words.

4.3 Counting

Our ACA features include the following counting features: count of words with length > 15 characters and number of words repeated in page >= 10 times. In addition, the CA features [3] include the following counting features: number of words in page, number of words in title, average length of words, amount of anchor text, fraction of visible content, and compressibility [3].

4.4. Cosine Similarity

Our ACA features introduce a group of cosine similarity-based features. The group contains the following features: cosine similarity between document body with title, document body with keywords, document body with description, title with document body, title with keywords, title with description, keywords with document body, keywords with title, keywords with description, description with document body, description with keywords, and description with title.

We compare words in the first component with words in the second component. For example, if we have five words in title represented by a vector \( x = \{1,1,1,1,1\} \), then the document body will be represented by the vector \( y = \{1,3,2,1,0\} \) if the second word in title is repeated three times and the fifth word not found in document body. We use the following formula for cosine similarity.

\[
\text{CosSim} (X, Y) = \frac{\sum_{i=1}^{n} (X_i \cdot Y_i)}{\sqrt{\sum_{i=1}^{n} X_i^2} \cdot \sqrt{\sum_{i=1}^{n} Y_i^2}}
\]

where the \( X \) vector represents words in the first component, \( Y \) represents words in the second component, \( n \) the number of words, and \( \cdot \) is the inner product of two vectors.

4.5 N-Grams Syntax

The n-gram syntax [3] is used to judge the correct grammar and strength of coherence of words with each other. In our work, we use independent n-gram likelihoods and conditional n-gram likelihoods for \( n = 2, 3 \) and 4. The independent n-gram likelihood of a document is defined as follows [3]:

\[
L_{n-grams}(doc) = \prod_{i=1}^{n} L_{n-gram}(w_i | w_{i-1})
\]

where \( w_i \) is the \( i \)-th word in the document, \( L_{n-gram}(w_i | w_{i-1}) \) is the likelihood of the \( i \)-th word given the \( i-1 \)-th word, and \( n \) is the order of the n-gram.
IndepLH = 1/K \sum_{i=0}^{k-1} \log P(w_i + 1, ..., w_i + n)

where IndepLH is the independent n-gram likelihoods, K the number of n-grams, P the probability of ith n-gram, w_i is the ith word in an n-gram, and n the number of words in an n-gram.

Documents with a high IndepLH value are composed of infrequently occurring n-grams. The conditional n-gram likelihood metric of a document is defined as follows [3]:

CondLH = 1/K \sum_{i=0}^{k-1} \log P(w_n|w_{i+1}...w_{i+n-1})

where CondLH is the conditional n-gram likelihood, p(w_n|w_{i+1}...w_{i+n-1}) is the probability of ith n-gram with a condition that n-1-gram occur

5. EXPERIMENTAL RESULTS

We conducted many experiments to measure the accuracy of detecting Arabic spam pages by CA features, our ACA features, and ALL features (the combined CA and ACA features). We used different train-test split percentages and three classifiers: Decision Tree, Naïve Bayes, and LogitBoost [33].

We measured the F-measure, which is a measure of accuracy [29]. It considers both the precision \( p \) (or specificity) and the recall \( r \) (or sensitivity) of the test to compute the score. \( p \) is the number of true positives (correct results) divided by the number of true positives plus the number of false positives (all returned results) [30]. \( r \) is the number of true positives (correct results) divided by the number of true positives plus the number of false negatives (results that should have been returned) [31]. The equations of F–measure, precision, and recall are:

\[
\text{Precision } (p) = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positives}}
\]

\[
\text{Recall } (r) = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false negatives}}
\]

\[
\text{F-measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

In all our experiments, we repeated the experiment ten times, and we report averages with 95% confidence interval as error bars in the Figures.

Fig. 3 shows an example of training set size for CA features, ACA features, and ALL features. For CA features with Decision Tree J48 [33], the highest F-measure value was 0.95 when the training percentage was 66%. For ACA features with Decision Tree J48, the highest F-measure value was 0.96 when the training percentage was 66%. For ALL features with Decision Tree J48, the highest F-measure value was 0.98 when the training percentage was 66%.

![Fig. 3. Results using different training set sizes for ALL, CA, and ACA features by Decision Tree.](image)

![Fig. 4. Results using different training set sizes for ALL features, CA features, and ACA features by Naïve Bayes.](image)
When Naïve Bayes and LogitBoost were used as the classifier, the results were as shown in Fig. 4 and Fig. 5, respectively.

In the next experiment, we divided ALL features into four groups, namely globally popular words, counting, cosine similarity, and n-grams syntax.

As shown in Fig. 6, for globally popular words with Decision Tree J48, the highest F-measure value was 0.6 when the training percentage was 66%. For counting, with Decision Tree J48, the highest F-measure value was 0.96 when the training percentage was 66%. For cosine similarity with Decision Tree J48, the highest F-measure value was 0.9 when the training percentage was 66%. For N-gram with Decision Tree J48 the highest F-measure value was 0.91 when the training percentage was 66%.

We then ranked the features using the gain ratio attribute from attribute evaluator with 10-fold cross-validation [33] to separate ALL features into significant features and weak features with respect to effect on classification accuracy. As shown in Table 2, the best feature was cosine similarity of description with keywords. The weakest features (not shown) was count of words with length >15 characters.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cosine Similarity of Description with Keywords</td>
</tr>
<tr>
<td>2</td>
<td>Fraction of One Hundred Globally Popular Words in Page</td>
</tr>
<tr>
<td>3</td>
<td>Fraction of Page Words Drawn from One Hundred Globally Popular Words</td>
</tr>
<tr>
<td>4</td>
<td>Cosine Similarity of Description with Document Body</td>
</tr>
<tr>
<td>5</td>
<td>Fraction of Fifty Globally Popular Words in page</td>
</tr>
</tbody>
</table>

Table 2. Top 5 Ranked features amongst ALL features.

Summary. ALL features performed better than CA and our ACA features alone. The highest F-measure value was 0.98. Decision tree was the best classifier for ALL features, CA features, and ACA features. The counting feature group performed better than the other groups when we used Decision Tree classifier, but the cosine similarity group was the best of the when we used Naïve Bayes classifier. N-gram group was the worse.

6. CONCLUSIONS AND FUTURE WORK

We have proposed a number of features for detecting Arabic Web pages that contain spam. We have also collected a corpus of Arabic Web pages (spam and non-spam) manually using search engines such as Google, Bing, AltaVista, Maktoob, Ayna. We used ALL features (combined CA and ACA features) to classify Arabic Web pages into spam and non-spam. We experimented with three classifiers. The best results we obtained were with combined CA and ACA features using the Decision Tree classifier. The highest F-measure value was 0.98.

Possible future work is to use natural language processing techniques to recognize artificial text, use more ACA features, collect larger Arabic corpus, and combine more classifier techniques.
REFERENCES


[22] G. Stringhini et al, Detecting Spam mers on Social Networks in ACSAC ’10 Dec. 6-10, 2010, Austin, Texas USA.


