Web Mining Techniques to Block Spam Web Sites

Esraa M. El-Mohdy
Computer Science Department
Faculty of Specific Education
Mansoura University

A. F. El-Gamel
Computer Science Department
Faculty of Specific Education
Mansoura University

Hanan E. Abdelkader
Computer Science Department
Faculty of Specific Education
Mansoura University

ABSTRACT
The aim of this paper is to introduce a system based on web mining techniques to prevent spamming web pages. The system relies on content analysis, used features are Uniform Resource Locator(URL), Number of words in page Title, Globally Popular Keywords(GPK) and N-GRAM. The proposed system used Decision Tree(DT) rules ; which is the best classifier to detect Web spam content. It produces accuracy of .97 % in detecting spam web sites.

Keywords
Web Mining, Spam Web Sites ,Decision Tree.

1. INTRODUCTION
With the great growing of the World Wide Web(WWW), there is a massive amount of web pages mankinded on every imaginable human rights to daily News for sports as described in online news articles, forums, and blogs. These pages may contain also a mix of computer data like graphics, videos, voices, multimedia, and pictures. The web includes a large number of users from different geographic regions. Users go to search engines such as Yahoo , Bing and Google for useful information. It may be recovered by millions of web pages for each search request , but only searches for a few selected web pages [1].

Given the amazing amount of information that can be obtained on internet, users usually specify beneficial web sites by requesting a search engine. At the request of the search determines the relevant search engine on the web site and displays users links to these sites, usually in batches of 10-20 link [2].

Search engine spam is an undesirable site that receives a lot of revenue from processing the content and links of a web site. People who spam search engines are called spammers or spammy content for search engines [3].

Spam is any deferred action just to promote a website's web page in search engine results, commensurate with the true value of the page. Web Spam is a web page that is the result of spam. Spam on the web is intentional doctrinaire of search engines indexes. It is one of the methods of search engines optimization. Implementing spam content on the search engine reduces unwanted and excessive results[4].

To determine the most useful information among the countless web pages available, users firstly depend on search engine. Search engines usually classify a huge number of web pages and provide pages which appear more related to user queries ranked by popularity and relevance. Users usually visit higher-rated web pages and ignore other pages [5].

The purpose of this Processing is to make their pages more suitable for user requests, thus doctrinaire search engines to raise the rank of a spam website to be included in top ten links that appear on the front page of Search Engines Results Page (SERP) [1].

The site contains malicious software that automatically installs itself on the system when the site is opened. The site can also affect the financial situation by continuing to own information such as bank account number, password ,and other financial information, and the Internet spam can be very serious from the user point of view. Since a spam site can attack the victim's system in different ways [4].

2. RELATED WORK
The literature contains numerous papers on the subject of Spam Web sites ,where the subject is examined from several points of view. This section displays little of these papers that are linked to paper topic: Detecting spam on the Internet detecting Arabic and non Arabic web based spam, and dedicated studies to assess relation between spam and popularity.

Mohammed A. Saleh , et.al had represented improvement of Arabic spam web pages detection using new robust features . They have suggested unprecedented collection of features which mends the detection of spammy Arabic web sites. These Features contain: Globally Popular Keywords (GPK), Sentence Level Frequent Words (SLFW), and Character N-Gram Graph (CNGG) features. They referred to new proposed set of features as features B in contrast to the state-of-art featured which referred by features A. they have combined their (B) features with the state-of-art (A) features to get (AB) features and then fed the resulting AB features into different classification algorithms include Ensemble Boosting with Bagging and Decision Tree ensemble methods, Random Forest classifiers, and Decision Tree J48 to obtain their results. In their results they obtained an F-measure of about 99.54% with the Random Forest classifier. They applied their new features on a dataset of about 15962 Arabic web pages which containing spam and non spam sites. they also compared their results with results of a previous studies in the field of Arabic spam web pages and they found that, their results (Fmeasure of 99.64%) have exceeded all their results (98%) with the same dataset they used in their study (Dataset 2010).[1]

Alexandros Ntoulas, et.al had represented detecting spam web pages through content analysis. They go on their realization of web spam : injecting theatrical generated pages into the web to effectiveness search engine results, for driving traffic to assured sites for profit or fun. This study looks at some of the techniques that are not previously described to automatically detect spam pages, and examine the validation of these techniques in segregation and when assembled using classification algorithms such as Amount of anchor text. Number of words in the page title. Average length of words . Number of words in the page, Fraction of visible content, Compressibility and Independent n-gram likelihoods. their inferences correctly recognize 2,037 (86.2%) of the 2,364 spam
sites (13.8%) in their arbitrated set of 17,168 sites. While 526 spam and non spam sites were misidentified (3.1%) [2].

Sumit Sahu, et. al had represented a web spam detection using different features. They discuss the features which are responsible for web page ranking. These Features include: N-gram feature and Match Score feature. They also argue about the results of different classification techniques on their dataset which they are processing from the WEBSPAM-UK2006 Dataset. They also suggest a feature that helps detect spam on the web [4].

Maria Soledad Pera had represented a structural, content-similarity measure for detecting spam documents on the Web. They provided a new approach to identify spam web sites. By looking at the coding of web sites and the content, they improve a spam detection tool which is inexpensive computational factors. Word correlation used to analyses the content is precompiled, and reliable, because they can exactly disclose 84.5% of legitimate/spam web sites. They realized that SPAM detection methods outweigh current Spam approaches by at least 3% in terms of F-measure [6].

Mohammed N. Al-Kabi et, al had represented an online Arabic web spam detection system, their study is an enhancement of past work in this area. It was included implementation and the design of system for detecting an online Arabic Web spam. Depending on mathematical foundations and algorithms, which can detect the Arabic link and content of web spam based on the tree of the spam detection conditions, be based on feedback used by a custom Web browser. Users can take a decision on any web page, through their feedback, so that they can decide if the web sites in the browser are pertinent to certain queries or not. Proposed system used linking features and content separated from Arabic web sites to mark whenever each web page will be classified as spam or non spam. The system also tried to benefit from user feedback to improve its execution automatically [7].

Heider A. Wahsheh, et al had presented A link and content hybrid approach for Arabic web spam detection. This study was a resurgence to a series of spam Arabic studies on the web produced by the authors. The paper was consecrate for building the first Web spam detection system on Arabic content/link. This system was qualified for extracting a set of link and content of Web sites features, and building considerable Arabic Web spam dataset. The set of data generated consist of three sets with following percent of the contents of the spam: 2%, 30%, and 40%. These groups have been collected with different percentages of spam content by the crawler included in proposed system. Automatic classification used depends on features in standard data set. Proposed system used Decision Tree rules; which considered the best classifier which can detect spam Arabic content/link. Proposed system helped clean up the SERP program from all URLs of spam web sites in Arabic. It produced accuracy of 93.1034% for Arabic link based, 90.1099% for Arabic content based, and 89.011% in Arabic content detection and link Web spam. Depending on the collected data set and perform the analysis [8].

Mohammed N. Al-Kabi, et al, had provided an determining of spam effect on Arabic web pages popularity. In this search. Many top rated web pages were selected to determine potential spam behavior on the web. Web pages used spam techniques for boosting their rates inside Search Engines. Results Page (SERP) like Page Rank Algorithm. Term Frequency-Inverse Document Frequency (TFIDF), and Hyperlink-Induced Topic Search (HITS) Algorithm. The results of this study showed that some of these popular sites use techniques that are considered unwanted techniques according to the instructions for search engine optimization [9].

Tarek Amr Abdallah et.al. had presented a URL-based web page classification: with n-gram language models, they had been recently suggested a new method to classify web pages URL. N-gram language model was presented to this task as a way for providing scalability and competitive accuracy in larger data sets their way allowed to classify new URL addresses with an invisible subclass. In this paper they extended their display and contained extra results to support the proposed approach. They explained the parameters associated with the n-gram language model and test their impact on the produced models. Their results showed that their method was competitive in terms of accuracy with the best known methods but also scaled well for larger datasets [10].

Radwan Jaramh, et al, had represented a Detecting Arabic Spam Web Pages Using Content analysis. In this paper, They proposed unprecedented features to enhance the ranking of Arabic websites in spam and non spam under several classification algorithms such as: Naïve Bayes, LogitBoost, and Decision Tree. they compared their features, which they called Arabic Content Analysis (ACA) features, to the latest Content Analysis (CA) features to detect spam in English Web. They showed that increasing CA features with their ACA features increased the accuracy of detection of Arabic spam sites compared to CA features only. When collective, ACA with CA features accurately characterized 5,536 pages of the 5,645 Arabic spam pages which they used to test with a FP rate of 1.9% by using DT classifier. furthermore, they characterized topmost ranked features with Gain Ratio method [11].

3. PROPOSED SYSTEM

The main challenges that meet researchers in the field of detecting spam web pages are the lack of dataset. There is no public and standard dataset published over the internet for this type of problem, so the dataset was collected manually. 7000 web pages were collected from April 2016 to December 2017. Some of the spam web pages were collected using search engines (Yahoo, Google, Bing) with a spamming query such as pornography contents, pages support terrorism at this link https://afteegypt.org/blocked-websites-list-ar and pages spreading false news, and violence.

The collection of non-spam pages are collected from a lot of trusted sites, such as governmental, educational, news sites and human rights.

Some examples of spam sites in collected database are presented in table1, also the reason for considering the website as spam:

<table>
<thead>
<tr>
<th>Table1: Sample of Arabic and English spam web pages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arabic spam sites</strong></td>
</tr>
<tr>
<td><a href="http://www.aljazeera.com/">http://www.aljazeera.com/</a></td>
</tr>
<tr>
<td><a href="http://carvideoTube.com/ar/search?q=arab+sex+web">http://carvideoTube.com/ar/search?q=arab+sex+web</a></td>
</tr>
<tr>
<td><a href="https://www.madamasar.com/newshttps://arabsex.info/tag/arab-sex">https://www.madamasar.com/newshttps://arabsex.info/tag/arab-sex</a></td>
</tr>
<tr>
<td><a href="https://abaysexy.com/">https://abaysexy.com/</a></td>
</tr>
<tr>
<td><strong>English spam sites</strong></td>
</tr>
<tr>
<td><a href="http://www.xnxx.com">http://www.xnxx.com</a></td>
</tr>
</tbody>
</table>
Many features have relied on a collected corpus of Spam web pages such as: URL of the page, Number of words in page title, Globally Popular Words and N-gram.

In the following part the proposed system features will be illustrated:

3.1 URL of The Page

A URL (Uniform Resource Locator), as the name indicates, provides a way to determine a resource on the web, the hypertext system that works over the internet. URL includes the name of the protocol to be used to access the resource and a resource name. First part of the URL defines which protocol to use. Second part defines domain name or IP address where the resource is situated.

In the URL test, the URL was compared to the white list that contains allowed sites database, and if it was found there, then the site will be opened without testing its content. And if it was not found there, another test will be done to detect its existence in the unwanted sites in the black database, and if it was found, then the site will be blocked directly, and if it was not found, then its content will be checked.

Some examples of the sites (spam, nonspam) in the database are presented in Table 2:

<table>
<thead>
<tr>
<th>Popular Keywords</th>
<th>English Keyboard</th>
<th>Arabic keyboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youtube</td>
<td>D,j,d,f</td>
<td>نغمات</td>
</tr>
<tr>
<td>Facebook</td>
<td>Facebook</td>
<td>توهاخله</td>
</tr>
<tr>
<td>Visisk</td>
<td>Tdsf;</td>
<td>ملحمه</td>
</tr>
<tr>
<td>Google</td>
<td>Google</td>
<td>موحن</td>
</tr>
</tbody>
</table>

3.3 Globally Popular Keywords

Globally popular Keywords are those being searched widely using search engines. Lots of websites take usefulness of these globally popular keywords to trick page ranking algorithms of the search engine to get the highest rankings in the first web page of the search results. Oftentimes, these keywords are used in spam Web sites for attracting users of search engine and achieved a large number of visitors, which translated into celebrity or money [11].

Globally Popular Keywords computed as follows: Count of words with length more than 15 characters and number of repeated words in page more than or equal 10 times, if it found spam web sites[1].

Some examples of the Globally popular keywords are presented in Table 3:

Table 3: Sample of globally Popular keywords

3.4 N-GRAM

N-gram is used for a variety of different tasks. For example, when developing a language model, n-gram is used to develop not only unigram models but also the bigram and Trigram models. Microsoft and Google have developed web-based n-gram models that can be used in a lot of tasks such as text summarization, word breaking and spelling correction.

N-gram are widely used in natural language processing tasks and text mining. They are basically a set of words that occur in a particular frame and when you calculate n grams you normally move one word forward (although you can move X words forward in more advanced scenarios).

Here N can be 1- unigram, 2 - bigram, 3-trigram and so on, X=Num of words in a given sentence K, the number of n-grams for sentence K would be:

\[ Ngrams_K = X - (N - 1) \]

3.4.1 Unigram

A matrix of unilateral words was extracted using unigram where every element in the matrix of words out of this equation

\[ Ngrams_K = X - (N - 1) \] where N=1[17], was compared to the elements exists in the database of unilateral spam words in the data collection. The results of this comparison were subjected to test, so if the number of words is greater than the threshold value which is greater than or equal 3 as a result of a test on a set of sites then the site will be blocked.

By testing the number of Unilateral words that express spam words within the database on a sample of different sites, it turns out that the spam sites contain 3 words or more, while the normal sites are less than 3.

Table 4 show Sample of checked sites and number of Unilateral spam words in these sites as shown:

3.2 Number of words in Page Title

Count the number of words in the Title tag of the page. IN normal Page titles are in the average five to six words long. If number of words in the title more than or equal 24 [2] it is considered as a spam web page and added to black-list database.
Table 4: Sample of checked Sites and number of Unilateral spam words in sites

<table>
<thead>
<tr>
<th>Checked URL</th>
<th>No. of words</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>porntubesdb.com</td>
<td>17</td>
<td>Spam</td>
</tr>
<tr>
<td>porntab.co</td>
<td>5</td>
<td>Spam</td>
</tr>
<tr>
<td>anybunny.com</td>
<td>2</td>
<td>Non-spam</td>
</tr>
<tr>
<td>tubevintageporn.com</td>
<td>16</td>
<td>Spam</td>
</tr>
<tr>
<td>spankbang.com</td>
<td>18</td>
<td>Spam</td>
</tr>
<tr>
<td><a href="http://www.marieclaire.com">www.marieclaire.com</a></td>
<td>2</td>
<td>Non-spam</td>
</tr>
<tr>
<td>saa-recovery.org</td>
<td>1</td>
<td>Non-spam</td>
</tr>
<tr>
<td><a href="http://www.dailymail.co.uk">www.dailymail.co.uk</a></td>
<td>30</td>
<td>Spam</td>
</tr>
<tr>
<td>/<a href="https://www.redtube.com">https://www.redtube.com</a></td>
<td>14</td>
<td>Spam</td>
</tr>
</tbody>
</table>

3.4.2 Bigram
A matrix of binary words was extracted using $\text{Ngrams} = X - (N - 1)$ where $N=2$ \([17]\), every element in the matrix was compared with the binary spam database of binary spam words exists in the database. The results of this comparison were subjected to test, so if the number of words is greater than the threshold value which is greater than or equal 3 as a result of a test on a set of sites then the site will be blocked.

By testing the number of binary words that express spam words within the database on a sample of different sites, it turns out that the spam sites contain 3 binary words or more, while the normal sites are less than 3 binary words, Table 5 show Sample of checked Sites and number of binary words.

Table 5: Sample of checked sites and No of binary spam words in sites

<table>
<thead>
<tr>
<th>Checked URL</th>
<th>No. of words</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.streampornmovies.com">www.streampornmovies.com</a></td>
<td>2</td>
<td>Non-spam</td>
</tr>
<tr>
<td><a href="http://www.yubuporn.com">www.yubuporn.com</a></td>
<td>3</td>
<td>Spam</td>
</tr>
<tr>
<td>/<a href="http://sexetc.org">http://sexetc.org</a></td>
<td>3</td>
<td>Spam</td>
</tr>
<tr>
<td>www2.fuckmymelons.com</td>
<td>3</td>
<td>Spam</td>
</tr>
<tr>
<td><a href="http://www.xxxproposal.com">www.xxxproposal.com</a></td>
<td>4</td>
<td>Spam</td>
</tr>
<tr>
<td><a href="http://www.youjizz.com">www.youjizz.com</a></td>
<td>1</td>
<td>Non-spam</td>
</tr>
<tr>
<td><a href="http://www.shepherd.net">www.shepherd.net</a></td>
<td>2</td>
<td>Non-spam</td>
</tr>
<tr>
<td>orangeporn.info</td>
<td>4</td>
<td>Spam</td>
</tr>
<tr>
<td>eskimotube.com</td>
<td>10</td>
<td>Spam</td>
</tr>
</tbody>
</table>

4. FRAMEWORK:
In the following figure, the steps of the proposed system are explained.

Fig.1: Block Diagram for test web site
Pseudo code to detect spam web sites According to the title is shown in The following part :

Start
Read Title
Count Words
IF count number >= 24 then
Spam URL
End If
End

Pseudo code to detect spam web sites according to the body is shown in The following part :

Start
Read body
Count words of GPK-length and repeated GPK
IF GPK-length > 15 and repeated GPK >= 10 Then
Spam URL
End If
IF number of n-gram(unigram-bigram) With spam words > 3 Then
Spam URL
End If
Open URL
End

5. DECISION TREE CLASSIFICATION

Decision Tree (DT) is one of the common classification techniques. It is presented as a graph of decisions which consists 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 [8].

The finding of a Target with the help of decision trees starts by preparing a set of solved cases. The whole set is then divided into a training set, which is used for the induction of a decision tree, and a testing set, which is used to check the accuracy of an obtained Target [16]. Attributes defining each case are described input data (unigram, bigram, GPK, Title) and among them one attribute is selected that represents a decision for the given problem Target (Spam, Nonspam).

Table 6 shows a training set. Each object's value of each attribute is shown, together with the class of the object (Target).

Table 6. Spam websites Dataset sample

<table>
<thead>
<tr>
<th>URL</th>
<th>bigram</th>
<th>GPK</th>
<th>Unigram</th>
<th>title</th>
<th>target</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.freepornx.org">www.freepornx.org</a></td>
<td>large</td>
<td>high</td>
<td>large</td>
<td>high</td>
<td>spam</td>
</tr>
<tr>
<td>sexuhot.com</td>
<td>large</td>
<td>low</td>
<td>large</td>
<td>low</td>
<td>spam</td>
</tr>
</tbody>
</table>

Decision tree is obtained using the Entropy Scale, which is a statistical method to measure the value of information in different fields and their effect on the target field. The decision attribute is calculated what is known as the rate of gain:

\[
\text{Entropy}(S) = -D_{\text{spam}} \log_2 D_{\text{spam}} - D_{\text{nonspam}} \log_2 D_{\text{nonspam}}
\]

\[
D_{\text{spam}} = \frac{\text{No of examples have "spam" decision}}{\text{Total No of Examples}}
\]

\[
D_{\text{nonspam}} = \frac{\text{No of examples have "nonspam" decision}}{\text{Total No of Examples}}
\]

In the following part, Figure 2 shows part of The Decision Tree.
From the Decision tree shown in figure 5 rules can be:

1. If Bigram="small" and Unigram ="small" then Decision = "nonspam".
2. If Bigram ="small" and Unigram ="large" then Decision ="spam".
3. If Bigram="large" and GPK ="low" or "high" then Decision ="spam".

### 5.1 Evaluation of Decision Tree

Evaluation of the execution of a classification model is depending on the count of test records correctly and incorrectly by the model. This can be done using True positive (TP) which refers to spam sites that have already been blocked ,false negative(FN) which refers to spam sites that have not been blocked, true negative(TN) which refers to non-spam sites that have already been opened and false positive(FP) which refers to non-spam sites that have been blocked.

### Table 7. spam and non-spam web sites with true/false positive and true/false negative rate with decision tree

<table>
<thead>
<tr>
<th></th>
<th>BLOCK</th>
<th>OPEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPAM</td>
<td>.95</td>
<td>.04</td>
</tr>
<tr>
<td>true positive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False negative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NON-SPAM</td>
<td>.00</td>
<td>1.00</td>
</tr>
<tr>
<td>false positive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>true negative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Recall is the quotient of correctly specified items (TP) and all pertinent items (sum of TP and FN)[13].

Recall  = $\frac{TP}{TP + FN}$

Precision is the ratio of correctly specified items (TP) and all pertinent items (sum of TP and FP)[13].

Precision  = $\frac{TP}{TP + FP}$

F-measure is one of the most generally used single number measures in retrieving information, Machine Learning and Natural Language Processing, but it is depend on a mistake, and the defective presumptions make it unsuitable for using in most contexts[18].

F-measure = $2 \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$

Accuracy is the ratio of true results (both TP and TN) through the total number of cases studied [14].

Accuracy = $\frac{TN + TP}{TN + TP + FN + FP}$

Table 8 presents the results of applying the previous equation:

### Table 8. spam and non-spam web sites with true/false positive and true/false negative rate with decision tree

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.95</td>
<td>1.00</td>
<td>.97</td>
<td>.97</td>
</tr>
</tbody>
</table>

### 6. CONCLUSIONS

The main goal of this study is solving the problem of web spam detection. The proposed system is based on web mining. Features are extracted from the web through proposed system. These features are URL of the page, number of words in page Title, Globally popular Keywords and N-gram. Decision tree classifier is used to improve the Accuracy.

We plan to extend this work in the future to study and research the detection of the pornography pictures and videos in spammed Web sites and the detection of Malicious links that usually mix between Web security issues particularly malware types (Worms and Viruses) and Web spam techniques.

### 7. REFERENCES


