Web Mining Techniques to Block Spam Web Sites

Esraa M. EL-Mohdy Computer Science Department Faculty of Specific Education Mansoura University A. F. EI-Gamal 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 manned 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 higherrated 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 mend 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 (F-measure 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, contentsimilarity 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 International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 8, August 2018

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 impacted 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:

Arabic spam sites	Reason
http://www.aljazeera.com/	Support terrorism
http://carvideotube.com/ar/sea rch/?q=arab+sex+web	Pornography
https://www.madamasr.com/n ewshttps://arabsexi.info/tag/ar	spreading false news
ab-sex	Pornography
https://arabysexy.com/	Pornography
English spam sites	Reason
http://www.xnxx.com	Pornography

 Table1: Sample of Arabic and English spam web pages

http:// www.xvideos.com	Pornography
masralarabia.com	Terrorism
http:// xxxfuckporn.com	Pornography
http:// www.pornhub.com	Pornography

Many features have relied for 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 indicate, provides a way to determine a resource on the web, the hypertext system that works over the internet. URL includes the name of 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 :

Table2: sample of white-list and black-list database white-list database

http://www.touregypt.net/
http://wcm.portal.eun.eg:10040/wps/portal
https://www.egypt.gov.eg/arabic/info/default.aspx
http://www.iti.gov.eg/
http://www.asrt.sci.eg/
http://www.emigration.gov.eg/
http://www.arabicacademy.org.eg/
black-list database
www.yubuporn.com
www.yubuporn.com spankbang.com
www.yubuporn.com spankbang.com www.pornobanga.com
www.yubuporn.com spankbang.com www.pornobanga.com sergiooliveir.eu
www.yubuporn.com spankbang.com www.pornobanga.com sergiooliveir.eu www.bossvideotube.com
www.yubuporn.com spankbang.com www.pornobanga.com sergiooliveir.eu www.bossvideotube.com www.allsexclips.com
www.yubuporn.com spankbang.com www.pornobanga.com sergiooliveir.eu www.bossvideotube.com www.allsexclips.com videopornarchive.com

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

3.3 Globally Popular Keywords

Globally popular Keywords are those being searched Widely using search engines. lots of websites takes 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 :

Popular Keywords	English Keyboard	Arabic keyboard
يوتيوب	D,jd,f	يوتيوب
YouTube	YouTube	غخعفعلاث
Facebook	Facebook	بشؤثلاخخن
فيسبوك	Tdsf,;	فيسبوك
Google	Google	لخخامث
جوجل	[,[g	جوجل

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 the 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 : $Ngrams\kappa = X - (N-1)$

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 $Ngrams\kappa = X - (N-1)$ 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 :

Checked URL	No. of words	Туре
porntubesdb.com	17	Spam
porntab.co	5	Spam
anybunny.com	2	Non-spam
tubevintageporn.com	16	Spam
spankbang.com	18	Spam
www.marieclaire.com	2	Non-spam
saa-recovery.org	1	Non-spam
www.dailymail.co.uk	30	Spam
/https://www.redtube.com	14	Spam

 Table 4: Sample of checked Sites and number of Unilateral spam words in sites

3.4.2 Bigram

A matrix of binary words was extracted using $Ngrams\kappa = X - (N-1)$ 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

4. FRAMEWORK:

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

International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 8, August 2018

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

Checked URL	No. of words	Туре
www.streampornmovies.com	2	Non-spam
www.yubuporn.com	3	Spam
/http://sexetc.org	3	Spam
www2.fuckmymelons.com	3	Spam
www.xxxproposal.com	4	Spam
www.youjizz.com	1	Non-spam
www.shepherd.net	2	Non-spam
orangeporn.info	4	Spam
eskimotube.com	10	Spam



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

URL	bigra m	GPK	Unigra m	title	targe t
www.freep ornx.org	large	high	large	high	spam
sexuhot.co m	large	low	large	low	spam

International Journal of Computer Applications (0975 – 8887)
Volume 181 – No. 8, August 2018

		1011			8000 2010
www.heavy handfuls.co m	large	high	large	low	spam
www.spank wire.com	large	low	large	high	spam
www.daily mail.co.uk	large	low	large	low	spam
www.xtube. com	large	high	large	high	spam
www.freevi deoshere.co m	Large	low	large	low	spam
wcm.portal .eun.egl	Small	low	small	low	Non- spam
cu.edu.eg/a r/Home	Small	low	small	low	Non- spam
/www.arabi cacademy.o rg.eg/	small	low	small	low	Non- spam
http://www. asrt.sci.eg/	small	low	small	low	Non- spam

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 :

Entropy(S) =

 $-D_{spam}log2D_{spam}-D_{nonspam}log2D_{nonspam}$

 $D_{spam} =$

No of examples have "spam" decision Total No of Examples

D_{nonspam}=

No of examples have "nonspam" decision Total No of Examples

In the following part ,Figure 2 shows part of The Decision Tree



Fig.2: Decision Tree Diagram

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

	BLOCK	OPEN
SPAM	.95	.04
	true positive	False negative
NON-SPAM	.00	1.00
	false positive	true negative

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 International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 8, August 2018

the defective presumptions make it unsuitable for using in most contexts[18].

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

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
positiv	e a	nd tru	e/fals	e negative r	ate w	ith de	cision t	ree

Recall	Precision	Accuracy	F-Measure
.95	1.00	.97	.97

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

- Mohammed A. Saleh, Hesham N. El mahdy and Talal Saleh, 2014, "Improvement of Arabic Spam Web Pages Detection using New Robust Features", Journal of Computer Engineering, Vol.16, Issue 2,pp24-35.
- [2] Alexandros Ntoulas, Marc Najork, Mark Manasse and Dennis Fetterly, 2006, "Detecting Spam Web Pages through Content Analysis", the International World Wide Web Conference Committee, ACM 1-59593-323-9/06/0005, pp1-10.
- [3] D. Saraswathi and A. Vijaya ,2016, "Search Engine Spam Detection using an Integrated Hybrid Genetic Algorithm based Decision Tree", International Journal of Computer Applications, Volume 133 – No.10,pp 20-27
- [4] Sumit Sahu, Bharti Dongre and Rajesh Vadhwani, 2011," Web Spam Detection Using Different Features", International Journal of Soft Computing and Engineering, Volume-1, Issue-3, pp 70-73.
- [5] Apostolis Zarras, Antonis Papadogiannakis, Sotiris Ioannidis and Thorsten Holz ,2015," Revealing the Relationship Network Behind Link Spam ", online at https://www.ics.forth.gr/_publications/zarasPST2015
- [6] Maria Soledad Pera," A Structural, Content Similarity Measure for Detecting Spam Documents on the Web", https://pdfs.semanticscholar.org/.../2836c81e3c0d1802e76f 1acee604.
- [7] Mohammed N. Al-Kabi, Heider A. Wahsheh and Izzat M. Alsmadi, 2014, " An Online Arabic Web Spam Detection

System", International Journal of Advanced Computer Science and Applications, Vol. 5, No.2, pp 105-110.

- [8] Heider A. Wahsheh, Mohammed N. Al-Kabi and Izzat M. Alsmadi ,2013, " A link and Content Hybrid Approach for Arabic Web Spam Detection ", Intelligent Systems and Applications, Published Online December 2012 in MECS (http://www.mecs-press.org/), pp30-43.
- [9] Mohammed N. Al-Kabi, Izzat M. Alsmadi and Heider A. Wahsheh, 2015, " Evaluation of Spam Impact on Arabic Websites Popularity", Journal of King Saud University – Computer and Information Sciences, pp 222–229.
- [10] Tarek Amr Abdallah and Beatriz de La Iglesia ,2015, "URL-Based Web Page Classification: With n-Gram Language Models",Springer International Publishing Switzerland, CCIS 553, pp. 19–33.
- [11] R. Jaramh, T. Saleh, S. Khattab, and I. Farag,2011, "Detecting Arabic spam web pages using content analysis," International Journal of Reviews in Computing,vol.6, pp.1–8.
- [12] Meenakshi and Geetika, 2014, "Survey on Classification Methods using WEKA", International Journal of Computer Applications, vol. 86,no.18, pp. 16–19.

International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 8, August 2018

- [13] Boris Neubert, Sören Pirk, Oliver Deussen and Carsten Dachsbacher ,2010, "Precision and Recall as Appearance Space Quality Measure for Simplified Aggregate Details", Eurographics Symposium on Rendering.
- [14] Simone Bassis Anna Esposito Francesco Carlo Morabito and Eros Pasero,2015,"Advances in Neural Networks: Computational Intelligence for ICT", Springer, pp.219.
- [15] Powers, David M W ,2011, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation", Journal of Machine Learning Technologies. vol.2, pp. 37–63.
- [16] Vili Podgorelec, Ivan Rozman and Peter Kokol,2002," Decision Trees: An Overview and Their Use in Medicine", Journal of Medical Systems, DOI: 10.1023/A:1016409317640 ,pp.1-21.
- [17] David Sundby,2009, "Spelling correction using N-grams", http://fileadmin.cs.lth.se/cs/education/EDA171/Reports/20 09/david.pdf
- [18] David M.W. Powers,2014," What the F-measure doesn't measure",https://www.researchgate.net/publication/273761 233_What_the_F-measure_doesn%27t_measure.