Towards Building Arabic Web Spam Detection System

by:

Heider Ahmad Wahsheh

Supervisor:

Dr. Mohammed Naji Al-Kabi

Computer Information Systems Department

Yarmouk University

July 17, 2012

Towards Building Arabic Web Spam Detection System

by:

Heider Ahmad Wahsheh

B.Sc. Computer Information Systems, Yarmouk University, 2009

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Computer Information Systems in the Computer Information Systems Department, Yarmouk University, Irbid, Jordan

Approved by:

Dr. Mohammed N. Al-Kabi.....**Chairman**

Assistant Professor of Computer Information Systems, Yarmouk University.

Dr. Bilal M. Abu-A'ata.....**Member**

Assistant Professor of Computer Information Systems, Yarmouk University.

Dr. Amer F. Al-Badarneh.....**Member**

Associate Professor, Department of Computer Information Systems, Jordan University of Science and Technology.

July 17, 2012

ACKNOWLEDGMENT

I would like to thank God for giving me the patience to work hard and overcome all the research obstacles.

My full gratitude to Dr. Mohammed Al-Kabi for his supervision, and precious advice. Without his support this work would not have been possible.

My deep thanks go to my thesis committee: Dr. Bilal Abu-A'ata, and Dr. Amer Al-Badarneh for their willingness to give of their time and experience.

I would like to extend my thanks to Dr. Izzat Alsmadi for his valuable guidance and help, and Dr. Ahmad Saifan who was so flexible and kind to me.

My thanks to my friends for their honest friendship, care, and for being kind to provide help and support.

Appreciation to my family, which has been always by my side, my father and mother for their motivating love and care. My brothers and sisters: Majd, Reem, Mohammad, Abeer, Ghadeer, and Yarub whom I have always had near me.

Heider Wahsheh

July 17, 2012

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

FP	**False Positive**
GP	**Genetic Programming**
HFSSL	**Harmonic Functions based Semi-Supervised Learning**
HITS	**Hyperlink-Induced Topic Search**
IDR	**Internal link Death Rate**
IGR	**Internal link Growth Rate**
IPR	**Internet Penetration Rate**
KL	**Kullback-Leibler**
KS	**Kappa Statistic**
LM	**Language-Model**
MAE	**Mean Absolute Error**
PPC	**Pay Per Click**
RAE	**Root Absolute Error**
RMSE	**Root Mean Squared Error**
ROC	**Receiver Operating Characteristic**
RRSE	**Root Relative Squared Error**
SEM	**Search Engine Marketing**
SEO	**Search Engine Optimization**
SERP	**Search Engine Results Pages**
TF-IDF	**Term Frequency-Inverse Document Frequency**
TP	**True Positive**
WITCH	**Web spam Identification Through Content and Hyperlinks**

ABSTRACT

Web is dynamic, each day the number of Web pages added to the Internet is increasing, and this relatively affects the total volume of the Internet. Web Search engines represent a major outlet to access the Internet world, information, or documents. The main goal of these search engines is to display the largest percentage of URLs of relevant Web pages inside Search Engine Results Page (SERP). Two main characteristics that distinguish good search engines are the speed and effectiveness, in addition to comprehensiveness which represents the coverage of new added materials to the Internet. Many Web sites owners act as spammers and try to mislead the search engines by using illegal Search Engine Optimizations (SEO) tips to increase the rank of their Web pages. This causes gaining more visitors for marketing and commercial goals. This study is dedicated to build the first Arabic Web spam detection system, which is capable to extract the set of content and link features of Web pages, in order to build an Arabic Web spam dataset. The constructed dataset contains three groups with three percentages of spam contents: 2%, 30%, and 40%. These three groups were collected through the embedded crawler in the developed system. The developed system used the rules of Decision Tree; as the candidate classifier to detect Arabic Web spam. The developed system produced the solutions of Arabic Web spam detection to improve Internet Arabic content value to the users. This system helped to clean the SERP from all URLs referring to Arabic spam Web pages. It produced an accuracy of 90.1099% for Arabic content-based, 93.1034% for Arabic link-based, 94.1606% for Arabic cloaking, and 89.011% in detecting both Arabic content and link Web spam, based on the collected dataset and conducted analysis.

Key Words: Arabic Web spam, content-based detection, link-based detection, content/link Arabic Web spam, cloaking Web spam.

CHAPTER ONE

INTRODUCTION

Arabic language is the widely spoken Semitic language. The Arabic language is considered as a mother tongue of more than 350 million people who distributed mainly over twenty five Arabic countries. Arab countries occupies Middle East and North Africa (MENA) region, besides the Horn of Africa (Wikipedia.2012). Arabic language is widely used throughout the Muslim world; since the Arabic language is the language of the holy Quran (words of Allah) and prophet Mohammed Sayings. The Arabic alphabets consist of twenty eight letters written in a cursive way from right to left like Amharic the official language of Ethiopia and the second widely spoken Semitic languages, followed by Hebrew, followed by Tigrinya which is mainly spoken in Eritrea and Ethiopia, and Aramaic which mainly used by small portion of the population of the Middle East. Arabic alphabets are adopted by other languages such as Urdu, Persian, Pashto, etc. Arabic language uses special configurations with Arabic diacritical marks (Tashkil or Harakat) considered as vowels which may change the meaning of Arabic words. It is ranked as the fifth most spoken language world-wide, and it is one of the official languages used by the United Nations (UN) (Ryding.2005; Beseiso, et al, 2010).

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

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

this amazing world of information (Alrawabdeh.2009). This chapter aims to present preliminary explanations about Web spam problems. The sections of this chapter present the following:

- The first section presents an overview of the spam types. Showing Arabic usage percentages statistics on the Internet, besides showing preliminary reviews of some Web master tools to improve the results.
- The second section presents the idea of Web spam. Which presents many illegal techniques to increase the rank of the Web pages.
- The third section is dedicated to show the types of Web spam (content, link, and cloaking).
- The fourth section presents the idea of Arabic Web spam, and the difference between Arabic and general Web spam.
- The fifth section presents problem statement of this thesis.
- The sixth section presents the research objective of this thesis.
- The seventh section presents the thesis structure.

1.1 Overview

Many challenges and obstacles are emerging in the every day expanding Internet environment, whether for the Internet users or the Web sites owners. The Internet users need to retrieve the high quality relevant information which is relevant to their queries within a short period of time.

While the Web site owners aim in most cases to increase the rank of their Web pages within Search Engine Results Page (SERP) to attract more customers to their Web sites, and consequently gaining more visits, which in turn means more revenues (Alrawabdeh.2009).

Throughout the last decade, many illegal techniques have widely used. Three main spam techniques are presented as follows:

- E-mail spam.

E-mails: E- mails are considered as one of the most important forms of communications through the Internet; they are characterized by a set of properties such as simplicity, and free availability to all users through the Internet. These properties make it prone to illegal use in the form of spam.

Spam, or junk e-mails are un-wanted e-mails, sent to users without their knowledge. Such e-mails form one of the main problems in the Internet. Statistics estimate that the total number of e-mails sent daily is 14.5 billion, where 45% of these are spam e-mails (spamlaws.2012). These mainly are categorized into:

1. Advertising-related e-mails, with a percentage of 36% of all spam (spamlaws.2012).

2. Adult-related e-mails, making up around 31.7% of all spam (spamlaws.2012).

3. Financial-related e-mails, those comprise 26.5% of all spam (spamlaws.2012).

These spam e-mails have a negative impact on Internet users, and lead to a decrease in the public confidence, and a decrease in productivity and safety (spamlaws.2012).

Spam blockers are techniques that have been used successfully to block a lot of spam e-mails. Microsoft Network Company (MSN) blocks around 2.4 billion spam e-mails every day (Mohammad & Zitar, 2011; spamlaws.2012). Spammers developed techniques that avoid blockers, such as Image spam, which consists mainly of an image that is composed of characters and

symbols within the textual message, in order to avoid e-mail spam filters (Biggio, *et al*, 2011).

- Mobile phone spams

Users also suffer from Mobile phone spams, which refer to the unwanted Short Message Service (SMS). When users receive spam SMS's without their request. In general these SMS's contain commercial advertisements. The messages deceive the recipients. Unlike the e-mail spam, the spam SMS is not widely used, it is shorter than e-mail spam and has less structure (Yoon, *et al*, 2010).

- Web spam

Another illegal technique is Web spam or spamdexing, which aims to increase the rank of Web pages and Web documents by deceiving Web crawlers. The description of Web crawler is usually based on the manipulation of content and link features of the spammed Web documents (Dou, *et al*, 2010; Baeza-Yates & Ribeiro-Neto, 2010).

The Arab World constitute about 5% of the world population, only 3.3% of the total number of Internet users are Arab users and the Arabic content on the Internet is less than 1% of all available online content (Internet World Stats.2012a; Tarabaouni.2011). Expressed by Internet Penetration Rate (IPR), the usage of Internet in some of the Arab world countries is depicted in Figure 1.1 (Internet World Stats.2012b).

Yet, despite these statistics, the usage of Internet throughout the Arab world is witnessing a rapid increase every day, particularly in the fields of social networks, and e- commerce (Tarabaouni.2011).

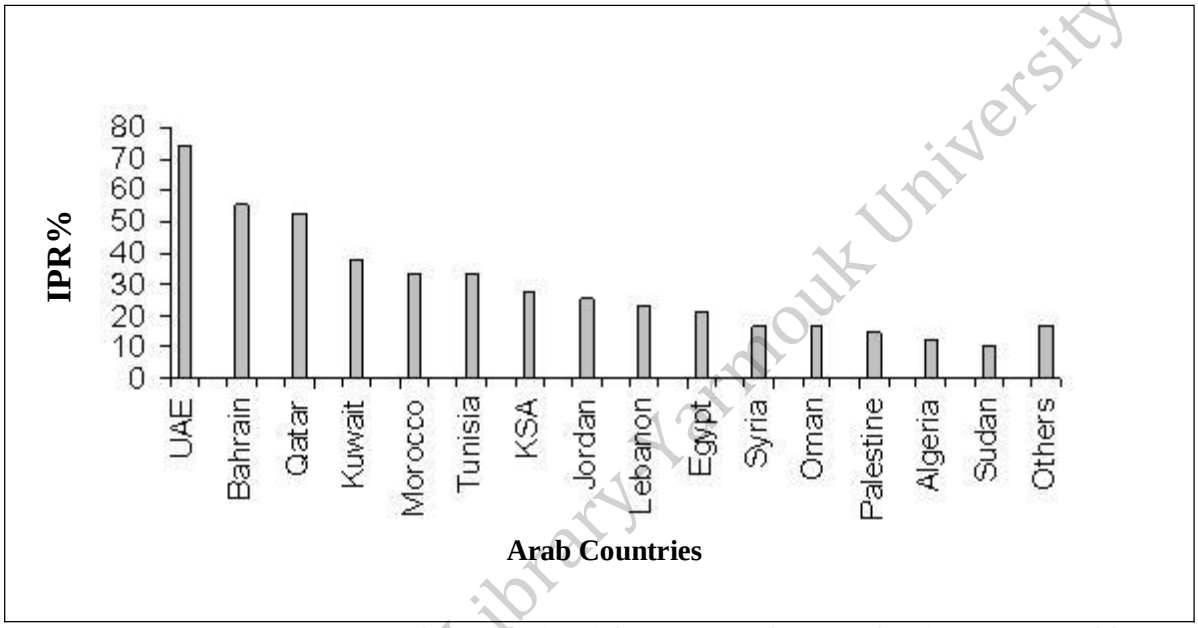


Figure 1.1: Internet Penetration Rate (IPR) in some Arab countries (Internet World Stats.2012b)

The statistics show that the United Arab Emirates, Bahrain, and Qatar are at the top of penetration rates list, while Iraq and Somalia are at the bottom of this list which are merged together under the label of Others in Figure 1.1. These differences in IPR can be explained by the regulatory, political environments, and the absence of mutual strategies to encourage the use of the Internet between the Arab countries (Tarabaouni.2011; Internet World Stats.2012).

Most Arabic Web pages are characterized by having unstructured format, lacking quality of the Arabic content and containing poor information, where the poor contribution appears within blogs and forums, which constitutes around 35% of total Arabic Web content (Tarabaouni.2011). The rest of Arabic content is distributed through e-commerce, newspapers, educational Web sites, and e-government Websites.

The free encyclopedia (Wikipedia) allows Internet users to publish and edit different articles in more than 200 natural languages (Almeida, *et al*, 2007), the contribution of Arabs is reflected by the percentage of the Web Arabic content which does not exceed 1% in best cases (Tarabaouni.2011).

Different search engines use different ranking algorithms which adopt many factors and metrics to manage the ranking process for different Web pages (Selvan, *et al,* 2012). These factors and metrics present both content and links features. Ranking algorithms represent a secret for different search engines; therefore these companies do not provide details about how they exactly rank the Web pages and consider these algorithms as their secrets that should not be known by other competitive search engine companies and Web spammers (Bendersk, *et al*, 2011). Web crawlers or robots constitute another part of a search engine; they are responsible for visiting different Web documents to be indexed (Batzios, *et al,* 2008).

Indexing is another essential part in search engines architecture which aims to help the search engine comprehensiveness and represents another secret for a search engine (Hochman.2012). Only the indexed Web pages are shown in the list of results, so the owners of Web sites try to improve their Web pages features to be rapidly indexed, which means increasing the chance for visibility in the top list results which known as search engine results pages (SERPs) (Hochman.2012; Dou, *et al*, 2010).

The owners of Web sites and the Webmasters of commercial Websites use Search Engine Optimization (SEO), Search Engine Marketing (SEM), to be more visible in SERPs, and the Banner advertisements, to attract user traffic, which increase company revenues (Dou, *et al*, 2010; Boone, *et al*, 2009).

SEO is based on the number of ethical methods and techniques aiming to reformat the content and material posted to the Internet. It helps the Web pages to meet the search engine requirements and gets a good rank to be considered as relevant Web pages in SERPs. SEO is applied to the content in the most essential HTML tags such as: <title>, Anchor text, URL, Headers tags (<h1>...<h6>), , and <Meta> tags. The improvement of the content of these tags will help the Web page to rank higher within SERPs (Boone, *et al*, 2009; Zhang, *et al*, 2005). SEM techniques also called pay per click (PPC) marketing are interested in optimizing the commercial Web pages. It helps the business growth, as it suggests the most popular marketing Keywords to appear inside the most important weighting tags in the Web pages. SEM is distinguished from SEO by its adopted technique which includes both pay per click (Adwords) and SEO (Dou, *et al*, 2010; Boone, *et al*, 2009; Zhang, *et al*, 2005).

Banner advertisements are constructed from the attractive elements like graphics, animations, flashes, sounds, and videos. Banners are usually linked to the company Web site advertisements. They seem more advantageous than SEO and SEM, because it is based on the idea of eye catching graphics which attract more users to click on the banners, and visit the Web pages advertisements (Boone, *et al*, 2009).

Some of Web site owners act as spammers or try to hire Web spammers, using the illegal SEO, SEM, and Banner advertisements methods and techniques in a complete or partial way to increase the rank of their Web pages. These methods define the term "Web spam" which fill the Internet with Web pages that deceive the search engines and take higher ranks than what they really deserve (Gyongyi & Garcia-Molin, 2005).

1.2 The Idea of Web spam

The concept "Web spam" refers to any illegal process aims to increase the rank of poor-quality Web pages and Web documents. This returns unrelated results to user's query (Gyongyi & Garcia-Molin, 2005).

Web spam uses many methods that manipulate the link and content features of the Web pages, such methods use:

- Invisible text: Web authors used the same color for the text and the background to hide the text manipulation of background colors in Web pages (Becchetti, *et al*, 2008; Ntoulas, *et al*, 2006; Gyongyi & Garcia-Molin, 2005).

- Keyword stuffing: This method is based on duplicating some Keywords or phrases, or inserting large number of unrelated words in the weighting tags of Web pages (Becchetti, *et al*, 2008; Ntoulas, *et al*, 2006)

- Tiny text: This method is based on inserting the keywords and phrases in a very small font and spread all over the Web pages, these tiny texts are not seen by the Internet users (Becchetti, *et al*, 2008).

- Internal links and External links: Illegal exchanges of the links cause a manipulation in the Web pages ranks such as: Article spinning, Scraper sites, and Doorway pages (Ermakova.2011; Becchetti, *et al*, 2008).

The goal of these techniques is to make the spam Web pages as normal Web pages, and to attract more Internet users to visit spam Web pages (Wang, *et al*, 2007a).

1.3 Types of Web spam

There are three main types of Web spam: content Web spam, link Web spam, and cloaking Web spam. Web spammers use these three types together in a single Web page or they can use any type they like (Gyongyi & Garcia-Molin, 2005).

- Content Web spam refers to any manipulation process that aims at changing the content of Web pages, by using some spamming techniques, such as: Keyword stuffing inside any of the following tags: the <body>, <title>, Headers <h1>...<h6>, , and <Meta> tags, beside using tiny and invisible text (Ntoulas, *et al*, 2006).

- Link spam is based on the manipulation of the link structure of Web pages, such as: inserting irrelevant links to point to other Web pages in illegal techniques.

  There are two main techniques used by link-based Web spam:

  1. Link hijacking is the most popular technique used for link-spam, which is based on using many links to point a target spam Web page called controlled page to arise its rank (Caverlee & Liu, 2007).

  2. The Honeypots is an indirect spam way, either entrapping a reputable Web page by inserting spam links, or using many links pointing to each other as controlled pages (Caverlee & Liu, 2007; Gyongyi & Garcia-Molin, 2005).

- Cloaking Web spam is based on the simple idea of producing two different versions for each spammed Web page; the difference comprises the content and quality. One of the two versions is meant to be with high

and valuable information quality, and is sent to the Web crawler to achieve a high rank. While the second version is meant to be a spam Web page, and is sent to the user browser (Lin.2009). Sometimes cloaking Web spam is viewed as a hybrid type of the two previous types (content and link spam), due to the similarity in the features with the content Web spam from one side, and the similarity with redirection which used in cloaking with the linked Web spam from the other side.

1.4 The Arabic Web spam

The Arabic Web spam is considered as a part of the general Web spam. Some common features such as: Number of words in the Web pages, number of words in the title, number of popular words in Web pages, number of internal, external and redirected links are traced in both the Arabic and the general Web spam (Wahsheh, *et al*, 2012a). Yet, the spammers in Arabic Web pages deal with some special considerations that are related in particular to the Arabic language model.

There are few datasets collections related to Arabic Web spam, hence the research in this field is still at an early stage. This is considered the main problem affecting the work of this thesis in that there is a lack in the references of Arabic Web spam Web pages. So we collected a large Arabic Web spam dataset, improve the Arabic Web spam features mentioned in the previous studies, (Wahsheh & Al-Kabi, 2011; Al-Kabi, *et al*, 2011; Jaramh, *et al*, 2011; Wahsheh, *et al*, 2012a; Wahsheh, *et al*, 2012b; Wahsheh, *et al*, 2012c; Al-Kabi, *et al*, 2012; Wahsheh, *et al*, 2012d), and built a content/link Arabic Web spam detection system.

1.5 Problem Statement

The top rank of the Web pages within SERPs, is very important to the e-commerce and commercial Web pages. The owners of Web sites can attract more visitors to their Web pages, and gain more revenue, through PPC when their pages appear in the top results of SERPs.

Some of the owners of Arabic Web sites use the spam techniques and methods, which violate SEO, SEM, and Banners advertisements in order to rate their Web pages higher than they deserve. Usually spammed Web pages are characterized by their low information quality, and the crowdedness of advertising content and links, which deceive the search engine and thus lead to irrelevant information that does not match users' queries.

The challenge of this research field is to find an optimal effective solution for Arabic Web spam problems.

1.6 Research Objective

This thesis aims to solve the Arabic Web spam detection problem. The key is to understand the methods and techniques which are used by spammers in Arabic spammed Web pages.

In this thesis we collected a large Arabic content/link based spam dataset than those collected in the previous Arabic Web spam studies. We adopted advanced content/link based features, such as: number of words in the Web page, number of words in the title, number of unique words in Web page, average words length, and number of internal, external and redirected links. The extracted features will be fed into classification algorithms, such as: Decision Tree, Logistic, and *K*-Nearest Neighbor (*K-NN*). The results of the classification algorithms are compared, and the best algorithm is

identified. Forwards, the rules of the best algorithm are implemented to build Arabic Web spam detection system.

Reducing the level of spam can help in saving time, effort and getting fast and relevant results that provide strong support to Arab Internet users, and allowing them to achieve the relevant results for their particular queries.

1.7 Thesis Structure

The following chapters of this thesis are organized as: Chapter two presents related studies to Web spam detection, chapter three presents the main weighting method and ranking algorithms for the Web pages, chapter four shows the research methodology, chapter five presents implementation and experimental results, chapter six describes the evaluation of our proposed system. Last but not least Chapter seven presents the conclusions and future work.

CHAPTER TWO

RELATED WORK

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

2.1 Non Arabic content-based Web spam detection

Ntoulas et al. (2006) use a various content-based features extracted from a real dataset of spam Web pages. They also used a number of heuristic methods for detecting content-based spam, and achieving high accuracy of detection using C4.5 classifier, which correctly identifies 86.2% of spam Web page within the dataset.

In their study Narisawa et al. (2007) propose unsupervised method to detect spam documents from a given collection of documents, by using the string equivalence relations. The unsupervised method presented as scalable and language independent on many Web documents in Japanese language.

Wang and Zeng (2007) produce a novel content-based trust model for Web spam detection, according to two real datasets one is in English and the other one in

Chinese languages. The results showed an enhancement on Web spam detection using SVM which yielded an accuracy of 90.13%.

Jone et al. (2007) explore measuring framework for poor quality search results caused by the Web spam problem. About 80 million Web pages from UK2007 WEBSPAM were indexed by one machine. The evaluation method presented a sensitive difference between baseline and filtered rankings.

In their study Benczur et al. (2007) propose a set of content-based features in which the occurrence of keywords play the main role in identifying the Web pages as spam or high value advertisement Web page. The experiment tests applied on the public known WEBSPAM-UK2006 dataset, and the results improved more than 3%.

A new method proposed by Goodstein and Vassilevska (2007) to detect the Web spam problem through two players game to identify the spam Web pages within search results. The novel game asks player to classify the Web pages as relevant, irrelevant, or passing to specific queries. This method was considered effective as truthfully voting Web spam algorithm.

Piskorsket et al. (2008) study the linguistic features, where more than 200 linguistic content-based attributes depending on two public available spam datasets (i.e. Webspam-Uk2006 and Webspam-Uk2007) where extracted to evaluate the proposed attributes.

Hayati and Potdar (2009) present the spam framework which focuses on the analysis of the current spam methods in Web 2.0 through two strategies: detection and prevention. The detection strategy was applied using content-based of the Blogs, comments, forums, opinions, online communities, Wikis and HTML tags. While the prevention strategy was based on the methods that prevent the spammers from

distributing the spam content on the Web server. The anti-spam framework showed that the detection methods consume the server-side resources, while the prevention strategy exploit and put the cost on the user side. So there is a need to improve robust prevention methods and do not increase user and server interaction complexity.

Pavlo and Dobro (2011) propose a new approach which based on the content analysis, that extract a new content-based diversity features depending on the frequency rank for the keywords and topics. The new features include both the linguistic features and statistical features. The well known WEBSPAM-UK2007 dataset was used as a training dataset. The conducted experimental results present the superiority of the diversity features to gain the high values for F-Measure in the range of 70-90 % in detecting content-based Web spam.

2.2 Non Arabic link-based Web spam detection

Gyongyi et al. (2004) propose a link-based algorithm called TrustRank which mainly based on forward links of the Web pages. It is assumed that the high reputable Web pages usually points to the good Web pages, and the unauthorized Web pages points to the spam Web pages. Gyongyi et al. (2004) collected manually a small number of reputable Web pages, and then used the link structure of the Web to find the other Web pages which likely considered as good Web pages. The results showed the effectiveness of the proposed TrustRank, which can detect the spam Web pages from a significant fraction of the web, based on a small number of good Web pages.

Many statistical analyses for large Web pages dataset was performed by Becchetti et al. (2006a) in order to detect Web spam using many link-based features such as: degree of correlations, number of neighbors, rank propagation through links, and TrustRank. They utilized link-based features to build many automatic Web spam

classifiers, and computing the score for each classifier, as well as computing the classifiers integration performance. The results of the proposed approach able to detect the link-based Web spam with an accuracy of 80.4%.

Chung et al. (2009) study special technique of link-based Web spam called hijacked links spam; which is based on bringing the rank scores from normal Web pages to the target Web spam pages. Chung et al. (2009) propose an algorithm for link hijacking detection, which is based on analyzing the features of the link structure which is neighbor to the hijacked Web sites. The results showed improvement in the accuracy of detecting the hijacked Web sites, where around 25% is over the other previous approaches.

Shen et al. (2006) study the link-based Web spam through using the link-based temporal information. Temporal features are divided into two groups; the first called Internal link Growth Rate (IGR), and the second called Internal link Death Rate (IDR). The first group (IGR) showed the ratio of the increased number of internal links in Web pages. While the second group presented the Internal link Death Rate (IDR), which define the ratio of the number of broken internal links to the number original internal links in the Web pages. The two groups can detect the spam behavior which try to add internal links to the pages to promote it by increasing the (IGR), and when the frequency of links changes in a spam Web page, which lead to increase the (IDR). The experimental tests used the SVM classifier to evaluate the proposed approach and achieved relatively higher percentage accuracy (40-60%).

The study of Becchetti et al. (2006b) focus on detecting link-based Web spam, and ignored the content-based Web spam features. They calculate the scores of the set of link-based features for each Web page, and applying the rank propagation and

probabilistic over the Web graph structure. They built the classifier which tested on the large Web link spam dataset. The tests used the ten-fold cross-validation, and the best classifier detect 80% of spam Web sites with only 2% as a false positives rate.

Many link-based Web spam has developed such as: Zhou et al. (2007) who develop a discrete analysis approach on the directed Web graphs for the analogue of classical regularization, which derived the powerful classification algorithms. The proposed approach applied on the real-world link spam detection and encouraging results have been obtained.

The study of Liang et al. (2007) presents another algorithm dedicated to the link spam detection, called R-spamRank. This algorithm produces an automated selection of spam Web pages especially those appears in the link farms. The authors manually collected a small spam dataset which considered as seeds for the evaluation process. They assigned spam values to the Web pages, and selected semi-automatically the most likelihood spam Web pages. The conducted test was based on a large dataset containing 5,000,000 Web pages, but only the top of 10,000 Web pages with the highest R-spamRank values were used. The results yielded an accuracy of 91.1% in detecting Web spam.

Caverlee and Liu (2007) analyze the page quality, and extract the link credibility through three distinct features. Semi-automatically technique is used to: Evaluate different pages link credibility, allow the personal user to assess the link credibility, and propose CredibleRank. The proposed CredibleRank algorithm is based on credibility metrics and quality of page scores superior to PageRank and TrustRank algorithms.

Geng et al. (2009) propose two link-based semi-supervised learning algorithms, to detect Web spam. The proposed algorithms merged the traditional Self-training with

the topological dependency based link learning. The results showed an increase in the performance of Web spam classifier, which allows feasible scheme for small Web spam samples detection.

Chung et al. (2010) use an online learning algorithm to monitor the host which can generate the link-based Web spam. Those researchers investigate on the Web spam seeds and extracted the set of link-based features which affect the PageRank score. The experiment tests used the archives of Japanese Web pages, and yields the precision between 56% and 73% and yields F-measure between 0.54 and 0.68.

The study of Niu et al. (2010) shows a new pattern using genetic programming (GP) which provides singularity functions to solve the link-based Web spam problem. The proposed pattern depends on analyzing the representation of link-based features of Web Pages through the public available dataset; WEBSPAM-UK2006, and the influence of the binary tree depth. The conducted tests showed that the following metrics (Recall, F-measure, and Accuracy) of new pattern are exceeded their counterparts of SVM by 26%, 11%, and 4% respectively.

Hayati et al. (2009) conduct a series of studies related to Web spam in Web 2.0 platforms, started with Hayati and Potdar (2009) in which the spammers insert their spam URLs in popular Web sites such as: social networking service, and Yahoo news. The spammers used spam methods such as: Web 2.0 spam which duplicates the amount of message spam, and Web spam bots which increase the spread of the spam content. Hayati and Potdar (2009) propose the Honey spam 2.0 tool which monitors the Web bot behavior. In their study Hayati et al. (2010a) continue the interesting on the spam bots, they used the action time and action frequency to detect the spam bots. The results showed that the accuracy enhanced and reaches 94.7%.

The study of Hayati et al. (2010b) explores an automated supervised machine learning technique to detect the spam bots inside 2.0 platforms. The new approach focuses on the navigation behavior, and compares between users and spam bots behavior. The conducted tests used Matthew Correlation Coefficient method, and have yields an accuracy of 96.24%.

West et al. (2011) build a link spam dataset which contains over 235,000 links of English Wikipedia, with extracted 40 features, by using Wiki metadata, landing site analysis, and external data sources. The conducted results showed enhancement in link-based Web spam detection.

Zhang et al. (2011) continue to work on the semi-supervised learning algorithm, by proposing a novel algorithm; called Harmonic Functions based Semi-Supervised Learning (HFSSL), where the labeled and unlabeled Web pages given weights based on the similarity in weighted Web graph. The results showed enhancement in the Web spam detection.

In their study Kumar et al. (2011) develop a new system called spamizer which able to detect the spam Web pages using content- based and link-based features. The spamizer analyzed several available link spam algorithms, such as: Relative spam Mass Estimation, Trustrank, Anti-Trustrank, Propagating Trust, and Distrust Scores and Reverse spam Rank. The experiments used the public known dataset WEBSPAM-UK2007, and they found that integrating the spamicity scores that generate from each algorithm increase the predictability for the spam and non spam Web pages.

2.3 Non Arabic content/link and cloaking Web spam detection

The study of Fetterly et al. (2004) focuses on identifying spam Web pages through the statistical analysis. Combination of content-based and link-based used to evaluate the three large datasets; of 150 million, 429 million, and 38 million redirected Web pages. The detection was based on the results of the statistical distribution, where the outliers indicate spam Web pages. While Fetterly et al. (2005) develop techniques dealt with two datasets; 151 million, and 96 million Web pages, to discover the phrase-level replication in spam Web pages. They used large number of machine-generated spam Web pages consisting of grammatically well-formed German sentences.

Wu and Davison (2005) report cloaking and redirection techniques as important spamdexing techniques. It produces a realistic view of content-based and link-based methods to detect cloaking and link redirections, through the computation of three different copies for each Web page. The analysis results estimated (through two used datasets) that 3% of the first dataset and 9% of the second dataset used the cloaking technique to increase the rank of their Web pages.

A Quantitative Study of forum spamming which uses a context-based and reported by (Niu, *et al,* 2006). The importance of the spam forums and splogs is due to three main perspectives: search users, spammers, and forum sites. Niu et al. (2006) in their study focus on the content-based and cloaking spam, and showed that the spam forums were used extensively. The spam forums supported by popular forum programs (which able the spam forums to occupy the top 20 search results for most popular keywords). The spam comments also used to increase the traffic on the honey spam forums. The results of splogs showed that more than half blogs are spam. The

researchers proposed context-based analysis; based on the cloaking analysis, to automatically detect spam.

Another study dealt with the blogs by Kolari et al. (2006) present the use of content-based and link-based spam features in blogs. The researches present the SVM model on content-based and link-based to detect Web spam, and achieved good results for recognizing splogs.

In their study Chellapilla and Chickering (2006) show the ratio of cloaking SERP which based on the query properties such as: link popularity and monetizability. They proposed new metrics for detecting cloaked Web pages through normalizing the *TF* between multiple downloaded Web pages versions. The experiments claimed using 10,000 search queries and 3 million related SERP. The results presented that 73.1% of the cloaked popular search Web pages are spam, and around 98% of the cloaked monetizable Web pages are spam.

The Svore et al. (2007) study the problem of Web spam, and proposed a new technique based on the rank-time features; which expressed both the content-based and link-based features. The proposed technique improves the performance of content-based Web spam classifier especially, when referring to the query-dependent rank-time features.

In their study Castillo et al. (2007) present a combination of link-based, and content-based spam detection features. The spam detection system divided into three phases; initially clusters the host graphs and labels all the hosts, then predicts labels to neighboring hosts, and finally uses the predicted labels as new features and retests the classifier. The researchers found that the linked hosts belong to the same cluster: either

both are spam or both are non spam. The results of the best classifier showed an accuracy of 88.4% with 6.3% false positives.

In their study, Geng et al. (2007) deal with the Web spam detection as a binary classification problem, using a new strategy called ensemble under-sampling classification. This strategy showed that the reputable Web pages are easy to be found than the spam Web pages. The public dataset WEBSPAM-UK2006 was used, which contains over than 8,000 spam hosts, and applied three learning algorithms (C4.5, Bagging, and Adaboost). The experiment results showed that the adopted ensemble under-sampling classification strategy added enhancements to the Web spam detection performance.

Webb et al. (2007) claim that they have performed the larger characterization content-based and HTTP analysis methods on the 350,000 Web documents. The analysis of content showed the duplication of the information content and URLs redirections. The analysis of content-based spam Web pages divided the Web documents into five main categories: Advertising Farms, Parked Domains, Advertisements, Pornography, and Redirection. While the link-based analysis performed, showed that the spammers used narrow IP addresses ranges for the spam hosting.

Chau and Che (2007) propose machine learning approach which combines content-based and link-based features using feedforward and backpropagation neural network and SVM. The experimental results confirmed that the proposed machine learning approach exceeded the traditional approaches (i.e. keyword-based and lexicon-based approaches), but showed that the best accuracy in this study is driven from a limited dataset size.

Tian et al. (2007) propose new technique based on a new machine learning approach, which can detect both link-based features, and word-based features, which extracted using the combinatorial feature-fusion method. They used many human-engineered features constructed from the raw data, and used semi-supervised learning to classify the unlabelled test Web pages. The results showed the high effectiveness of semi-supervised learning, and the combinatorial feature-fusion method.

Becchetti et al. (2008) study several content-based and link-based Web spam techniques such as: keywords stuffing, links stuffing, redirection pages, duplicated content, and hiding text. Two well known datasets were used in the experiments UK2002; which contains 18.5 million Web pages and WEBSPAM-UK2006 which consists of 77.9 million Web pages. Their experiments results revealed that a number of techniques were able to detect around 88% of spam Web pages.

The URL redirection is considered as a one of the cloaking styles. In their study Vangapandu et al. (2009) report the employment of the redirection on the spam links. The study found that the redirection is widely used, about 40% of all links for different goals. Several JavaScript of URL, Meta and server sides' redirections were detected in the spam Web pages as internal and external links. The experiment results applied on the legitimate Alexa, UGA, and blogs datasets, showed that the percentage of using external redirection techniques by spammers is 46.81%, and the percentage of using external redirection techniques by legitimate sites is 53.19 %.

Chellapill and Maykov (2007) present the cloaking Web spam in form of spamdexing redirections with false content to Web crawler for indexing purposes, while the irrelevant content sent to user browser. They studied famous JavaScript redirection techniques, which are stronger than the static analysis and static feature based detection

systems. Their research found that the use of light weight JavaScript parsers is effective to predict the redirection of spam behaviors.

Erdelyi and Benczur (2011) propose a graph similarity based on the content and link temporal features which capture the linkage change of the neighborhoods hosts. The public spam dataset WEBSPAM-UK2007 used by the machine learning techniques and the conducted results showed an enhancement on the filtering mixed language domains, where the content-based features cannot be reliably used for classification.

A study by Abernethy et al. (2008) present a novel algorithm; called Web spam Identification Through Content and Hyperlinks (WITCH) which aims to learn the Web spam detection techniques on both Web sites and Web pages level. The Witch algorithm takes the advantages of the content-based features and the Web graph structure. The authors claimed that Witch algorithm outperformed the other algorithms in the scalable, efficiency, and the state-of-the-art accuracy using SVM which achieved accurate results in detecting Web spam.

Benczur et al. (2008) presents LiWA FP7 project for solving Web spam problem. The proposed project used the three main types of Web spam (content, link, and cloaking) on Web archive. The project architecture facilitates the relations between the Web archives in different hosts and countries. The project used the well known UK2007 WEBSPAM dataset which contains 100,000 Web pages. This architecture has yielded good results for Web spam detection.

Najadat and Hmeidi (2008) present a novel approach of Naïve Bayes which considered the users feedback. It is suitable to be used in the server-side to minimize the overhead with spam Web pages in Web servers. The experimental results yield the high Web spam detection accuracy percent with 80.2%.

Castillo et al. (2008) study the feedback of the users and converted it to the query log. For each user, a query log file was assigned. This log file contains: query words, document returned to the search engine, Web documents which the user is clicked on within clicked date and time, and the rank of retrieved documents. The researchers applied two approaches: Web spam detection, and query spam detection. Web spam detection removes spam link and content features from the query log graphs, while query spam detection eliminates all queries that gain a high number of spam Web pages.

Martinez-Romo (2009) proposes language model approach which extracted a combination of content-based and link-based features from two popular spam datasets (Webspam-Uk2006 and Webspam-Uk2007). Kullback-Leibler (KL) divergence was applied on the spam Web pages to characterize the relation between the two linked Web pages. The proposed model has improved the F-measure of Webspam-Uk2006, and Webspam-Uk2007 to about 6% and 2% respectively.

In their study Dai et al. (2009) use the historical information in the Web pages as a useful complementary factor to Web spam detection. The combination of current content, content temporal features, and some temporal link features were applied using supervised learning techniques. The results of their study help to improve the F-measure performance within the content features.

The study of Egele et al. (2009) present a novel approach to detect Web spam pages in SEPRs, based on the integration of both content-based and link-based features. Initially the proposed approach compute the importance for each Web page, depending on the heavy weighting tags with spam content, and the linked relations through link spam. Then the classifications techniques were used to distinguish the spam and non

spam Web pages. The J48 Decision Tree classifier used in the evaluation of their approach, and achieved false positive rate of 10.8%.

Lin (2009) presents the influence of cloaking techniques to increase the rank of Web pages. Lin proposed three techniques: TagDiff2, TagDiff3, and TagDiff4 to determine if the URLs are cloaked. The proposed techniques are based on the different changes in the HTML tag between the Web crawler and user browser. The conducted tests showed that tag-based methods exceed the link-based and content-based results in precision and recall. The Decision tree J48 uses the integration of content-based and tag features to yield an accuracy of 90.48%.

Araujo et al. (2010) propose a new methodology to detect spam Web pages based on the Qualified-Link (QL) analysis, and content-based features with the language-model (LM). Kullback-Leibler (KL) divergence was applied on the spam Web pages to find the relation between two linked Web pages based on both the content-based and link-based features. An automatic classifier was built to combine QL and LM features. The conducted results applied on WEBSPAM-UK2006, and WEBSPAM-UK2007 datasets, showed an accuracy of 89.4% and 54.2% respectively.

The studies of (Wang et al. 2007a; Wang et al. 2007b; Wang et al. 2010) focus on the information quality features in order to define the trust features of spam information, such as: currency, availability, information to noise ratio, authority, cohesiveness, and the popularity. Taken these features into consideration has shown an improvement in the accuracy of detecting content and link Web spam.

Spirin et al. (2011) survey different spam methods, and filtering algorithms. The existing solutions were dedicated to content-based, link-based, and non-traditional data (i.e. user behaviors, clicks, and HTTP sessions) Web spam detection. Their anti-spam

algorithms provided high successful Web spam detection results with an accuracy of 90%.

Saraswathi et al. (2011) study has focused on different methods for Web spam detection. A novel approach used machine learning to build Web spam detection tool. The UCINET software and SVM classifier were used to identify the spam Web pages, based on many proposed features such as: degree of centrality, links betweenness and Eigen vector value of the link, which identify the quantitative and qualitative link farm properties. Their proposed approach used the WordNet database through the semantic analyzer, and obtained useful information that successfully discovered the spam Web pages.

As a continuation to the studies related to blogs, Zhu et al. (2011) propose a framework for splog detection by monitoring the top-ranked results. The framework arranged the sequence of temporally queries and detected splogs based on the temporal behavior. The experiments showed a high accuracy on splogs detection.

Erdelyi et al. (2011a) present different categories for Web spam features based on recent advances in Web spam filtering. Three of machine learning algorithms (i.e. ensemble selection, LogitBoost and Random Forest) were used. The conducted tests were applied on the two well known available datasets WEBSPAM-UK2007 and the Discovery Challenge dataset DC2010. The tests used ensemble classifier to detect spammed Web pages, and the improvement results ranged between 5-7.5%.

2.4 Arabic content/link based Web spam detection

In their study, Wahsheh and Al-Kabi (2011) conducted series of studies dedicated to Arabic Web spam problem. In their study Wahsheh and Al-Kabi (2011) have manually collected s small Arabic Web spam dataset; containing around 400 Arabic content-based spam Web pages. Three classifiers were tested; Decision Tree, Naïve Bayes, and *K*-Nearest Neighbour (*K-NN*). The results showed that the *K-NN* has yielded a better accuracy than the other two classifiers in detecting Arabic Web spam pages. The study of Jaramh et al. (2011) follows the Wahsheh and Al-Kabi (2011) and proposed new content-based features to improve the Arabic Web spam detection. Their study applied three classifiers (Decision Tree, Naïve Bayes, and LogitBoost), and the results presented that the Decision Tree classifier achieved the best results.

Al-Kabi et al. (2011) have integrated the two previous studies (Wahsheh and Al-Kabi, 2011; Jaramh, et al. 2011), and proposed a set of new content-based features, and used a larger spam dataset than Wahsheh & Al-Kabi (2011). Three classification algorithms (Decision Tree, LogitBoost, and SVM) were used to detect Arabic Web spam. The results confirmed the superiority of the Decision Tree as the best classifier with an accuracy of 99.3462% to detect Arabic Web spam.

Wahsheh et al. (2012d) analyzes the behaviors of the spammers to create spammed Arabic Web pages. They computed the weights of the most ten popular Arabic words used in the content of the HTML tags, which used in the Arabic queries. The results present special key stuffing techniques used in the Arabic spammed Web pages. The conducted test used the Decision Tree classifier to evaluate the spammer's behavior, and achieved 90% accuracy to detect Arabic Web spam.

Al-Kabi et al. (2012) improve their previous studies on the content-based Arabic Web spam. They used a large Arabic content-based spam dataset which contains 15,000 Web pages, that were collected by a special crawler. These Web pages were identified manually as spam or non spam. They applied four different classification algorithms (Naïve Bayes, Decision Tree, SVM, and *K-NN*) on the groups of the datasets, where the spam percentages were: 1%, 15%, and 50%. The results also revealed that the Decision Tree was the best classifier with 99.96% accuracy.

Machine learning is used to identify spam Web pages. Wahsheh et al. (2012a) conducted a study based on the machine learning algorithm to identify the content-based Arabic spam Web pages. The spam dataset was collected from three resources: the first is Extended-Arabic-2011 Web spam dataset, and the second is UK-2011 spam dataset where they were built by (Wahsheh, *et al*, 2012a). The third is a portion of the WEBSPAM-UK2007 spam dataset. Experiments were based on two algorithms (Naïve Bayes, and Decision Tree). The conducted tests of the proposed features show high accuracy results with Decision Tree which is better than Naïve Bayes in detecting Arabic spam pages, and yields acceptance results in detecting non Arabic Web spam.

All the previous Arabic Web spam studies (Wahsheh & Al-Kabi, 2011; Al-Kabi, et al.2011; Jaramh, et al. 2011; Wahsheh, et al. 2012a; Wahsheh, et al. 2012b; Wahsheh, et al. 2012c; Al-Kabi, et al. 2012; Wahsheh, et al. 2012d) tried to identify the best classification algorithm for the content-based Arabic Web spam detection, which almost unanimously that indicate the Decision Tree classifier is the best. Therefore Wahsheh et al. (2012b) based on the 15,000 Arabic spam Web pages, enhanced more content-based features, and built the pioneering Arabic Web spam detection system using the rules of Decision Tree classification algorithm. The experiment results presented an accuracy of 83% using the proposed system.

In an attempt to solve the problem of the Arabic link-based Web spam, Wahsheh et al. (2012c) studied the link-based spamming technique which is used by Arabic Web spammers. Wahsheh et al. (2012c) present that the spammers used the link-based spam techniques in the Arabic Web pages. The first Arabic link-based spam Web pages dataset was built by them. Many link-based features were extracted, and two classifiers (Decision Tree, and Naïve Bayes) were applied to evaluate the Arabic link-based Web spam. The conducted experiment show that spammers use a link spam farms technique between Arabic spam Web pages. The results of Decision Tree yield the best accuracy of 91.4706% to detect link-based spam Web pages.

CHAPTER THREE

WEIGHTING METHOD AND RANKING ALGORITHMS

This chapter discusses several weighting methods and ranking algorithms which are used by search engines, in order to rank the Web pages within search engine results pages (SERP). In this chapter we discuss the three main types of these algorithms; Term Frequency-Inverse Document Frequency, PageRank, and Hyperlink-Induced Topic Search. The spammers attempt to violate these three algorithms to gain the best possible rank for the spammed Web pages in the SERP.

3.1 The Term Frequency-Inverse Document Frequency (TF-IDF)

The Term Frequency-Inverse Document Frequency (*TF-IDF*) is a weighted schema which shows the importance of the words in the document (Gadg, *et al,* 2011). The value of *TF-IDF* of each term is dependent on the frequency of that term beside the number of documents which has that term, and the occurrence of the terms inside the Web pages (Gadg, *et al*, 2011). The terms which appears in special positions in the Web page such as: the <body> tag, Anchor text, URL, Headers (<h1>...<h6> tags), <meta> tags, and within the Web page <title> present more important than the other terms in the rest of Web page positions (Liu, *et al*, 2007).

Baeza-Yates and Ribeiro-Neto (2010) present four formulas for Term Weighting; $F_{i,j}$, *TF*, *IDF*, and *TF-IDF* as shown in the following mathematical equations:

Let,

k_i is an index term and d_j is a document.

$V = \{k_1, k_2, ..., k_t\}$ is the set of all index terms.

$(w_{i,j} \geq 0)$ is the weight associated with (k_i, d_j).

The weights $w_{i,j}$ are computed using the frequencies of occurrence of the terms within documents. $f_{i,j}$ is the frequency of occurrence of index term k_i in the document d_j. So the total frequency of occurrence $F_{i,j}$ of term k_i in the collection is defined as shown in formula (3.1):

$$F_{i,j} = \sum_{j=1}^{N} f_{i,j} \quad \text{.....(3.1)}$$

Where N is the number of documents in the collection.

Baeza-Yates and Ribeiro-Neto (2010) present the Luhn assumption which indicates that the weight of $w_{i,j}$ of index term k_i that occurs in the document d_j is relative to the term frequency $f_{i,j}$. This assumption means that increasing an occurrence of the term in the document, leads to get a highest weight.

The formula of Term Frequency *TF* is presented in formula (3.2):

$$TF_{i,j} = f_{i,j} \quad \text{.....(3.2)}$$

While the variant of *TF* weight is presented in formula (3.3):

$$TF_{i,j} = \begin{cases} 1 + \log_2 (f_{i.j}) & if\ (f_{i.j} > 0) \\ 0 & otherwise \end{cases} \quad \text{.....(3.3)}$$

The formula of Inverse Term Frequency *IDF* is presented in formula (3.4):

$$IDF_i = \log \frac{N}{n_i} \quad \text{.....(3.4)}$$

Where IDF_i is the i^{th} inverse document frequency of term k_i; n_i is the number of documents in which term i occurs at least once.

The best known term weighting schemes use combination weights of $TF_{i,j}$ and IDF_i factors.

The Term Frequency- Inverse Document Frequency formula is shown in the following formula:

$$w_{i,j} = \begin{cases} \left(1 + \log_2 \left(f_{i,j}\right)\right) \times \log_2 \left(\frac{N}{n_i}\right) & if \left(f_{i,j} > 0\right) \\ 0 & otherwise \end{cases} \quad \text{.....(3.5)}$$

Where $w_{i,j}$ is the term weight of the term k_i in the document d_j which refers to *TF IDF* weighting scheme; $f_{i,j}$ is the frequency of occurrence of index term k_i in the document d_j (Baeza-Yates & Ribeiro-Neto, 2010).

Spammers try to increase the *TF-IDF* scores in their spam content-based Web pages. They use many repeated and unrelated words in tags of an HTML such as: the <body> tag, Anchor text, URL, Headers (<h1>…<h6> tags), <meta> tags, and the Web page <title>, with many repeated and unrelated words in order to gain a higher *TF-IDF* score (Gyongyi & Garcia-Molin, 2005).

3.2 Hyperlink-Induced Topic Search (HITS) Algorithm

Hyperlink-Induced Topic Search (HITS) algorithm, also known as hubs and authorities, introduced by (Kleinber) in 1999, as a link analysis algorithm. It is proposed before the PageRank algorithm used for ranking Web pages (Selvan, *et al*, 2012).

HITS divided the Web pages into two main types: the first is called hubs; which indicate the Web pages that work as large directories, that not actually held the information, but it is points to many authoritative Web pages, which actually hold the information. So, a good hub represents a Web page that points to many other Web pages. The second type is called authority Web page which holds the actual information, and a good authority represents as a Web page that many hubs point to (Selvan, *et al*, 2012; Jayanthi, & Sasikal, 2012).

HITS computes two values for each Web page: the first value is for the authority which represents the score of the content-based Web page, and the second value is for the hub, which estimates the score for of its links to other Web pages (Selvan, *et al*, 2012)

Formula (3.6) presents the Authority Update Rule:

$\forall p$, we compute $A(p)$ to be:

$$A(p) = \sum_{i=1}^{n} H(i) \quad \text{.....(3.6)}$$

Where $A(p)$ is the Authority for p Web page; n is the total number of Web pages that linked to p; i is the Web page linked to p; and the $H(i)$ is the hub values for the i Web page that points to p (Selvan, *et al*, 2012).

Formula (3.7) explore the Hub Update Rule as shown below:

$\forall p$, we compute $H(p)$ to be:

$$H(p) = \sum_{i=1}^{n} A(i) \quad \text{.....(3.7)}$$

Where $H(p)$ is the Hub for p Web page; n is the total number of Web pages p connected to; i is a page which p connects to; and the $A(i)$ is the Authority values for i page (Selvan, *et al*, 2012).

The Web page classified as a good hub if it points to many good authoritative, and the Web page is good authority if it is indicated by many good hubs. The hubs values can be spammed through the link spam farms by adding the spam outgoing links to the reputable Web pages. The spammers target is to increase the authority values which violate the TrustRank algorithm. This means that spammers attempt to increase the hub values, and attract several incoming links from the spammed hubs to point to the target spam Web pages (Gyongyi & Garcia-Molin, 2005).

3.3 PageRank Algorithm

PageRank was proposed and developed by Google's founders (Larry Page and Sergey Brias) as a part of a research project about a new kind of search engines. It defines a numeric score which measures the degree of Web pages relevance to particular queries, and it is important due to the high score value of PageRank determines the list of SEPR for corresponding quires (Kerchove, *et al*, 2008).

PageRank can be seen as a model of user behavior. It assumes that there is a random Web surfer, starts from randomly Web page. Web surfers usually keep clicking on the forward links, and when the time passes they get bored and chose another random Web page. Therefore the PageRank computes the probability of Web surfer to randomly visit a Web page (Kang, *et al*, 2011).

The PageRank algorithm is considered as one of the main successful factors in Google. The last revealed algorithm from Google indicates that the PageRank algorithm is a link ranking one, which takes the number of internal links as an important factor in

page popularity. PageRank gives each page a score that determines the popularity of that page. The overall score of page p is determined by the importance (PageRank scores) of pages which have out links to that page p (Kang, *et al*, 2011). The generic formula which appears in literature for calculating PageRank score for a page p is shown in the following equation:

$$r(p) = \alpha \times \sum_{(q,p)} \frac{r(q)}{w(q)} + (1-\alpha) \times \frac{1}{N} \quad \text{.....(3.8)}$$

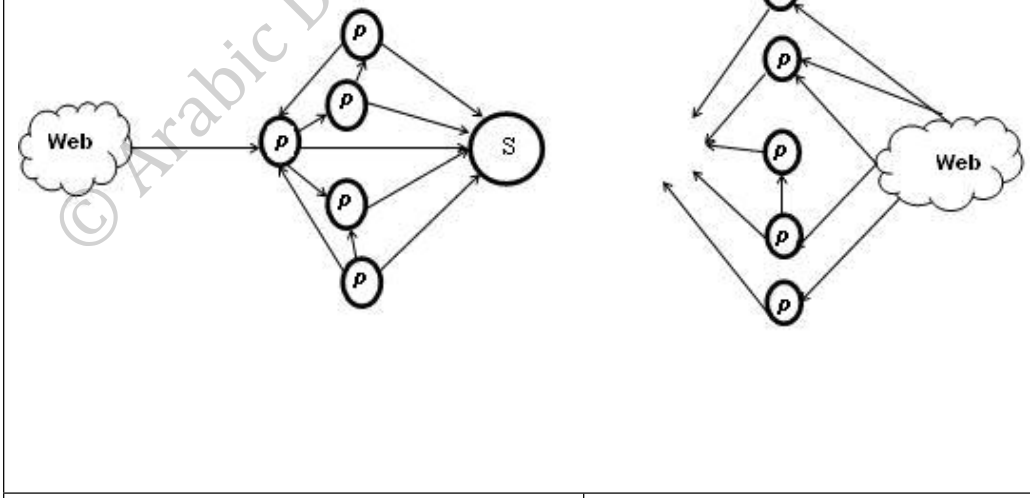
Where $r(p)$ is the PageRank value for a Web page p; $w(q)$ is the number of forward links on the page q; $r(q)$ is the PageRank of page q; N is the total number of Web pages in the Web; α is the damping factor which can be set between 0 and 1; (q, p) means that Web page q points to Web page p (Berlt, *et al*, 2010).

A Web page with a high PageRank score will appear at the top of the list of SEPR as a response to a particular query. Despite of this success for those search engines that use PageRank as ranking algorithm; spammers and malicious Web masters use some of PageRank algorithm problems to boost the rank of their Web pages illegally by using techniques that violates the SEO tips. Since PageRank is based on the link structure of the Web, it is therefore useful to understand how addition or deletion of hyperlinks influences it.

The degree of success in the link structure modifications is based on the degree of Web page accessibility by spammers. In most cases Web pages are inaccessible by spammer, so it is difficult for spammers to modify the link structures for such Web pages. Some Web pages on the other hand are partly accessible by spammers, hence, in a limited way spammers can post comments on such Web pages, such comments may

carry an external link from blog site to their spam page (Gyongyi & Garcia-Molin, 2005).

The third kind of Web pages that spammers have a full access on, are those Web pages owned by spammers. In such Web pages spammers try to create a link structure that work as a spam link farm, which is defined in Du et al. (2007) as heavily connected Web pages, created intentionally with the purpose of tricking a link-based ranking algorithm. In such case spammers will create a link structure that consists of a few boosting Web pages that may refers directly to each other and to the spam page in order to achieve some advantage by search engine ranking algorithms. Du et al. (2007) spammers can build different structures for a spam farm, such farm structure may be changed periodically in terms of the number of internal and external links, that is when spam filters drops spam links it is expected from the spammers to change their link structure by adding new links to their spam farm structure.



(a) Spam link farms structure (left) | (b) Normal link structure (right)

Figure 3.1: Two Main Web graph Structures (Du, *et al*, 2007)

Figure 3.1 shows a sample Web graph with two structures, the one on the left presents a set of densely connected Web pages (*p*), where each of which has links to

another as well to spam page which is the target whose rank is to be boosted. It appears in Figure 3.1a (left), which it has few links to the rest of the Web, and its goal is to boost the rank of spam Web pages by having too many internal links for its boosting neighbor's Web pages. On the other hand, Figure 3.1b (right) has a normal structure and consists of set of Web pages which have enough connections with the rest of Web graph. The differences between these two structures attract researchers' to study the properties of these two structures and the variations of the structure appear in the left Web graph (Du, *et al*, 2007).

It is known from the previous discussion that spammers have partial accessibility to some external Web pages that may have a good ranking score in search engines ranking vector. So it is expected from spammers to post links to those Web pages, because having a huge number of internal links on their spam page may achieve some improvement on its rank.

Figure 3.2 gives an example of a Web graph in which spammers make an attempt to boost the rank of spam page (*S*). The link structure used in Figure 3.2 is an example of optimal link spam farm used in (Gyongyi and Garcia-Molin, 2005; Largillier and Peyronnet, 2010) in which the authors proved how spammers can achieve benefit of having this structure. The structure consists of one target spam Web page (*S*). The spammers goal is to boost the PageRank of this target Webpage by pointing to page *S* using a set of Web pages $X = \{x_1, x_2, x_3\}$ in which the spammers have some accessibility (i.e. posting comments, adding links), spammers have also a full access on Web pages owned and created by them. So, the spammers also use their own set of Web pages $y = \{y_1, y_2\}$. This set of Web pages is used mainly to post links to the target page *S* in order to boost its rank. spammers will also add some external links from page *S* to the

Web pages $y=\{y_1,y_2\}$ however no out links will be posted on Web pages $y=\{y_1,y_2\}$, except those to the target page *S*.

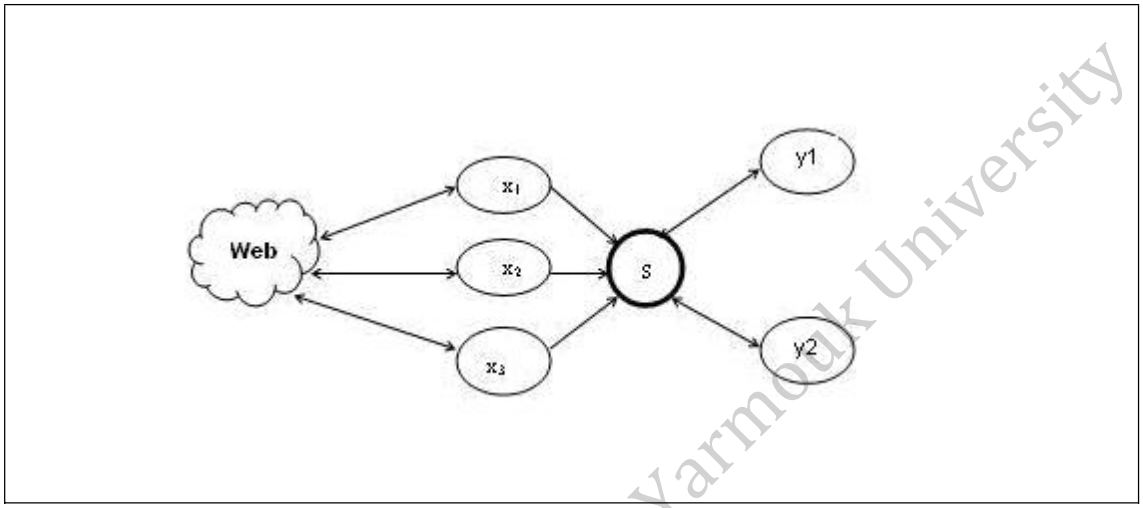


Figure 3.2: Optimal spam farm structure (Gyongyi & Garcia-Molin, 2005)

The total PageRank score of page *S* is maximized by the set of accessible (x_1, ...,x_3). The score that the target Web page gains from the boosting Web pages is calculated using the formula (3.9):

$$\sum_{i=1}^{3} \frac{r(p)}{out(x_i)} \quad \text{.....(3.9)}$$

Where *r*(*p*) is the PageRank; and *Out*(*x*) the number of accessible Web pages (Zhou, *et al*, 2009).

Every accessible page linked to the target spam page may have some contribution on its PageRank score. Such links are called hijacked links, (Du, *et al*, (2007). The total of PageRank scores of popular Web pages that have links (hijacked links) pointing to target spam Web pages is called leakage. The leakage gained by hijacked links is not known by spammers; however, their goal is to have as much hijacked links as it is possible.

The target page *S* PageRank score can be also maximized. If that page points to all Web pages created and maintained by spammers (boosting Web pages), given that those Web pages have no internal links except those from the *S*. So the search engine will, reach the spam farm through one of its hijacked links. It is possible then to crawl boosting Web pages through the external link from the target spam page (Chung, *et al*, 2009).

Finally, the *S* rank score can be also maximized if the set of owned Web pages $\{y_1, y_2\}$ have only external links to the target page *S*. This requires no links between boosting Web pages to each other. It requires also no hijacked links from outside world to the boosting Web pages (except from the *S*). The targeted page actually needs to point to all boosting Web pages to improve its PageRank score and to make every single Webpage in the whole spam farm accessible by search engines crawler (Du, *et al*, 2007).

CHAPTER FOUR

RESEARCH METHODOLOGY

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

1. Develop an embedded Web crawler; which is an automated tool, embedded in our new Arabic Web spam detection system. This crawler downloads the Web pages, parses all the hyperlinks, and the content in each Web page.

2. Build a large dataset of Arabic content/link spam Web pages relative to those built in the previous studies, which contains 28,000 Arabic Web pages, divided into two parts. The first part is called training dataset used to build Arabic content/link Web spam detection system, and contains 23,000 Arabic Web pages. While the second part is called test dataset, contains 5,000 Arabic Web page, and used to evaluate Arabic content/link Web spam detection system. The new dataset extended the last datasets used by (Wahsheh & Al-Kabi, 2011; Al-Kabi, *et al*,2011; Jaramh, *et al*, 2011; Wahsheh, *et al*, 2012a; Wahsheh, *et al*, 2012b; Wahsheh, *et al*, 2012c; Al-Kabi, *et al*, 2012; Wahsheh, *et al*, 2012d), using the enhanced Web crawler.

3. Develop a Web page analyzer to extract large number of features relative to those used in the previous studies. Categorized as three Web spam main types, for each Web page, as following: content-based features, link-based features, and cloaking features.

4. Use the three classification algorithms Decision Tree, Logistic Regression, and *K-NN* which are supported by Weka.

5. Compare the classification results of Decision Tree, Logistic Regression, and *K-NN* algorithms to identify the best classifier to detect Arabic spam Web pages.

6. Extract the rules of the best classification algorithm using the training dataset, to develop the decision maker as a final part of our Arabic content/link Web spam detection system.

7. Evaluate the Arabic content/link Web spam detection system, using the test dataset that contains 5,000 Web pages including spam and non spam.

Figure 4.1 summarizes the research methodology procedures.

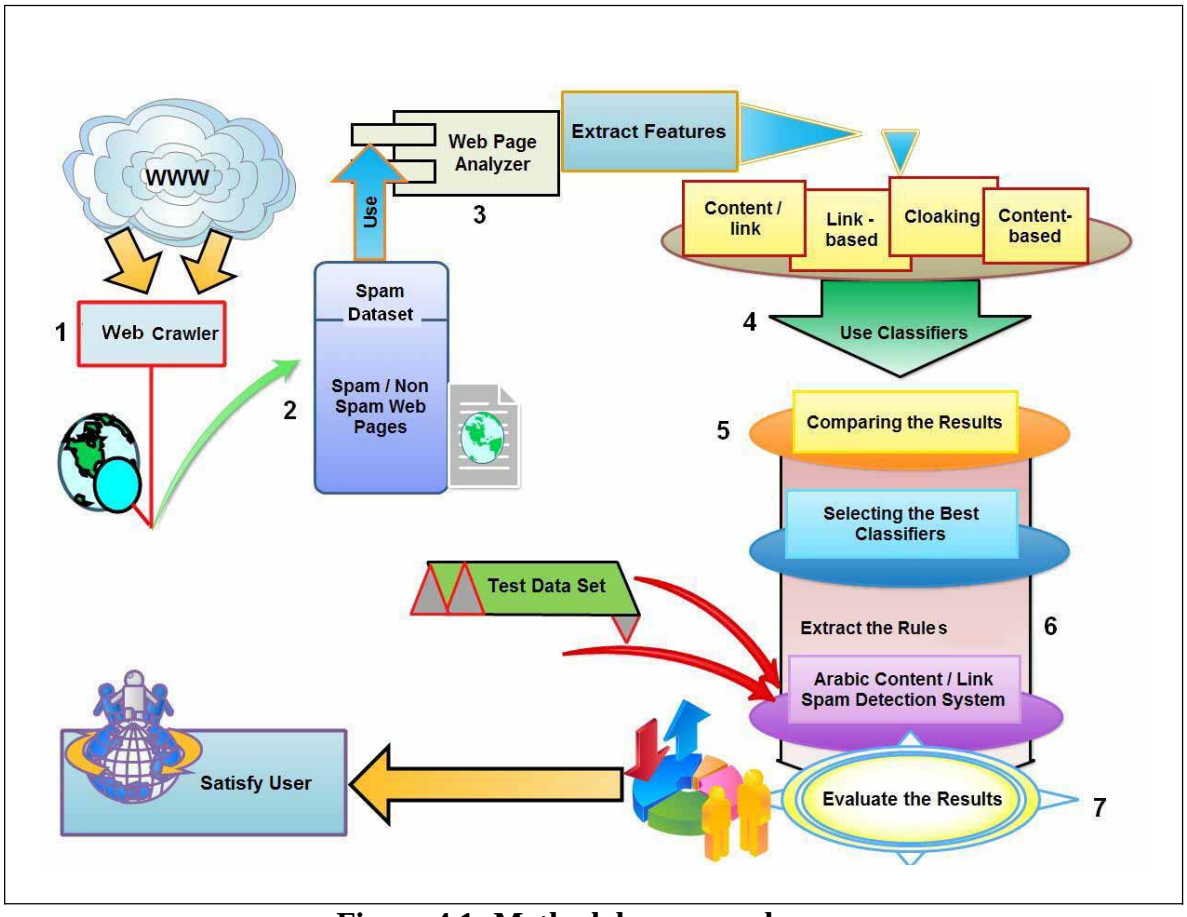


Figure 4.1: Methodology procedures

4.1 Develop an embedded Web crawler

Web crawlers are also known as spiders, or Web agents. Crawlers are types of software agents, which automate the traversing, fetching, sorting, and clustering Web pages, creates a copy of all Web pages before indexing. Crawlers traverse the Web by starting from a random Web page and continue by the following links to other Web pages. Sometimes crawlers need to exchange the information with other crawlers in order to notify their peers about sites with rich semantic content (Batzios, *et al*, 2008). So it appears as the main process in any search engine.

In this thesis we develop an embedded Web crawler in Arabic web spam detection system, which automatically parses through Web pages, downloads the Web pages, parsed all the Web pages elements; hyperlinks, and the content in each Web page.

4.2 Build an Arabic Web spam dataset

We have a lack of benchmark collection Arabic spam Web pages and this considered as one of the main challenges in Arabic Web spam detection process. In this thesis we built a large Arabic Web spam dataset containing 28,000 Web pages. Where 23,000 Web pages of the total Web spam dataset were used as training dataset which extended the datasets mentioned in (Wahsheh & Al-Kabi, 2011; Al-Kabi, *et al*,2011; Jaramh, *et al*, 2011; Wahsheh, *et al*, 2012a; Wahsheh, *et al*, 2012b; Wahsheh, *et al*, 2012c; Al-Kabi, *et al*, 2012; Wahsheh, *et al*, 2012d). The rest of the Web spam dataset consists of 5,000 Web pages was used as test dataset. The new dataset improves both the number of Arabic spam pages and their features as shown in the next sections.

The Web pages in the Arabic Web spam dataset are divided into two types: spam, and non spam Web pages. We split the spam training dataset into many groups based on the accuracy percentages of Arabic Web spam detection. Some groups got a

close accuracy values to each other; therefore we select the best percentage from those close accuracy percentages. The best three groups with different spam percentages were (2%, 30%, and 40%) of the dataset. Table 4.1 shows the Arabic spam dataset groups taxonomy.

Table 4.1. New Arabic spam dataset groups taxonomy

Close percentages values groups	Best percentage values of spam group	Number of spam Web pages	Number of non spam Web Pages
1%-15%	2%	460	22540
16%-32%	30%	6900	16100
33%-50%	40%	9200	13800

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

The spam Web pages consist of Arabic content-based blogs, forums, some of marketing Web pages, and advertising Web pages, which try to increase their possibility to appear at the top of SERP. The non spam Arabic Web pages can easily be found within the Web pages of universities and educational (.edu), ministries and governmental institutions (.gov), news sites (.com, .net) well known business companies (.com), and satellite channel (.tv).

Figure 4.2 shows an example of Arabic spam Web pages.

Figure 4.2: Arabic Web spam example

Figure 4.2 presents an example of spam Web pages, which duplicated some words as a Keyword stuffing methods to increase the *TF-IDF*, and used some spam techniques in the links to gain the highest possible rank in the SERP.

Figure 4.3 shows an example of Arabic non spam Web pages.

Figure 4.3: Arabic non spam Web page

Figure 4.3 presents an example of non spam Web pages; it is for the Yarmouk University; which is one of the governmental universities. It is considered as a trusted, and reputable Web page, and it is not used any illegal methods to increase its rank in the SEPR.

4.3 Develop Web page analyzer

We developed a Web page analyzer capable to extract the previous proposed features by (Wahsheh & Al-Kabi, 2011; Al-Kabi, *et al*,2011; Jaramh, *et al*, 2011; Wahsheh, *et al*, 2012a; Wahsheh, *et al*, 2012b; Wahsheh, *et al*, 2012c; Al-Kabi, *et al*, 2012; Wahsheh, *et al*, 2012d), and proposed new features. The new features which are

used in this study are divided into three groups: content-based features, link-based features, and cloaking features.

4.3.1 Content-based features

Spammers in Arabic Web pages used general and unique technique, with Keyword stuffing to increase the rank of their spam Web pages. The keyword stuffing spamming technique is based mainly on the duplication of some words in the main HTML elements. Spammers in Arabic Web pages used a unique keyword stuffing technique which is based on duplicating meaningless English words. Unique keyword stuffing technique is based on the relation between the meaningless English words and their corresponding Arabic letters that lie on the Arabic/English key board. (Wahsheh & Al-Kabi, 2011; Al-Kabi, *et al*,2011; Jaramh, *et al*, 2011; Wahsheh, *et al*, 2012a; Wahsheh, *et al*, 2012b; Wahsheh, *et al*, 2012c; Al-Kabi, *et al*, 2012; Wahsheh, *et al*, 2012d). This spam behavior leads to increase the rank of Arabic spammed Web pages, and from our point of view lead to deterioration of the quality of the Arabic content.

Figure 4.4 shows an example of Arabic spam Web page with general key stuffing technique, which duplicate the Arabic word 'chat' (دردشه، شات), many times without adding any meaningful information for the Web page.

Figure 4.4: Arabic spam Web page using keyword stuffing technique

Wahsheh et al. (2012d) studied the spam behavior within the top ten Arabic Keywords extracted by Google's free Search-Based Keyword tool (SBK), such as: 'chat' (دردشه، شات), 'Games' (ألعاب), 'YouTube' (يوتيوب), 'Facebook' (فيس بوك)'University' (جامعة), 'Forums' (منتديات), 'Songs' (أغاني), 'Photos' (صور),'Billiards' (بلياردو),'Trab' (طرب). Wahsheh et al. (2012d) found that the spammers used meaningless English words in Arabic spammed Web pages. The Latin letters of those meaningless English words lie on the same Arabic/English keyboard keys. Therefore the lengths of these meaningless English words always equal the lengths of Arabic words. Top ten Arabic words are not enough to detect the meaningless English words spammed technique. Thus, to address all the different topics that spammers might seek to use it in their techniques. Our Web page analyzer convert every English word to its corresponding meaningful/meaningless Arabic word, by convert every Latin letter of

the English word to its Arabic letter sharing the same key on the Arabic/English keyboard. Then the analyzer use the database which contains Arabic word list which developed by (Attia.2011). The database of Arabic word list contains nine million Arabic words. So the analyzer check the availability of every converted Arabic word in the database, to determine whether it is meaningful or not. Our analyzer considered the converted Arabic word a spammed word if it is found in the database. spammed Arabic words are generally meaningless.

Figure 4.5 shows an example of Arabic spam Web page using meaningless English words Keyword stuffing technique. The Arabic words are not included in the top ten Arabic Keywords used in the search engines, but the spammers used them to increase the rank of the Arabic Web page, and it is reducing the quality of the Arabic Web pages.

Figure 4.5: Arabic spam Web page using meaningless English words Keyword stuffing technique

Web page analyzer computes the content-based features which mentioned in the Arabic literature (Wahsheh & Al-Kabi, 2011; Al-Kabi, *et al*,2011; Jaramh, *et al*, 2011; Wahsheh, *et al*, 2012a; Wahsheh, *et al*, 2012b; Wahsheh, *et al*, 2012c; Al-Kabi, *et al*, 2012; Wahsheh, *et al*, 2012d). The Web page analyzer extracts the following set of new content-based features:

1. The number of meaningless English/Arabic words in the <body> elements of Web pages, where the spammers used this technique widely.
2. The total number of characters in all <Meta> elements of a Web page under consideration.

3. The number of characters in the <Meta> element of the Web page. So if we have *n* as a number of <Meta> elements in specific Web page, the Web page analyzer compute the number of characters in each <Meta> as a independent element of other <Meta> elements in this specific Web page.
4. The total number of Arabic/English words in the all <Meta> elements in a specific Web page.
5. The total number of Arabic/English words in each <Meta> elements in a specific Web page.
6. The total number of English words in all <Meta> elements inside a specific Web page.
7. The total number of Arabic words in all <Meta> elements inside a specific Web page.
8. The symbol words or characters are composed of letters, unique characters, and punctuation marks that may appear in some Arabic Web pages. Therefore it is considered as a candidate to be one of Arabic spammed features. So we extract the total number of characters of the symbols in all <Meta> elements of the Web page. Our Arabic content/link Web spam detection system will check if the spammers use those strange symbols to increase the rank of their Web pages, or not.
9. The total number of characters of the symbols inside a specific Web page.
10. The total number of characters of the symbols in the <body> element.
11. The total number of Arabic characters in a specific Web page.
12. The total number of Arabic characters in the <body> element.
13. The total number of English characters in a specific Web page.
14. The total number of English characters in the <body> element.

15. The total number of words which composed by the symbols inside a specific Web page.
16. The minimum length of English word inside the <body> element, (we assume that the minimum word consist of three characters).
17. The minimum length of Arabic word inside the <body> element, (we assume that the minimum word consist of three characters).
18. The minimum length of Arabic/English word inside the <body> element, (we assume that the minimum word consist of three characters).
19. The minimum length of word which composed by the symbols inside the <body> element, (we assume that the minimum word consist of three characters).
20. The maximum length of Arabic/English word which composed by the symbols inside the <body> element.
21. The maximum length of Arabic word which composed by the symbols inside the <body> element.
22. The maximum length of English word which composed by the symbols inside the <body> element.
23. The maximum length of symbol word which composed by the symbols inside the <body> element.
24. The maximum length of Arabic/English word which composed by symbols inside a specific Web page.
25. The average lengths of words inside the <body> element.
26. The average lengths of Arabic words inside the <body> element.
27. The average lengths of English words inside the <body> element.
28. The average lengths of symbol words inside the <body> element.

29. The total number of <Meta> element inside a specific Web page.
30. The Web page size in Kilo bytes.
31. The total number of characters with the URL. Spammers try to insert many spam words in the spam URLs, to attract the users to use these URLs.
32. The complexity factor of Web page within lexical density in the <body> element.
33. The complexity factor of Web page within lexical density inside a specific Web page.
34. The total number of Arabic/English words inside the <title> element.
35. The total number of Arabic/English words inside the <body> element.
36. The total number of Arabic words inside the <title> element.
37. The total number of Arabic words inside the <body> element.
38. The total number of English words inside the <title> element.
39. The total number of English words inside <body> element.
40. The total number of symbol words inside the <title> element.
41. The total number of symbol words inside the <body> element.
42. The visible page fraction inside the <page> element.
43. The visible page fraction inside the <body> element.
44. The size of the hidden text inside a specific Web page. The spammers try to trick the search engines to see links and the content that are not visible to normal users. This can be done through embedding them in very small pictures, or using tiny text font, or using the same color as the page background.
45. The total size of compressed files inside a specific Web page.
46. The total size of compression ratio inside a specific Web page.

47. The total number of Arabic/English words without repetition inside the <body> element.
48. The total number of Arabic/English words without repetition inside a specific Web page.
49. The total number of Arabic words without repetition inside the <body> element.
50. The total number of Arabic words without repetition inside a specific Web page.
51. The total number of English words without repetition inside the <body> element.
52. The total number of English words without repetition inside a specific Web page.
53. The total number of symbol words without repetition inside the <body> element.
54. The total number of symbol words without repetition inside a specific Web page.
55. The total number of image/images inside a specific Web page. Spammers try to attract the users through using large number of meaningless images in their spam Web pages. They use these images to increase the traffic of visitors, so they can get more revenues.
56. The total number of links image/images inside a specific Web page.
57. The total number of the most popular Arabic words inside a specific Web page. We based on the most popular Arabic words that mentioned in (Wahsheh, *et al*, 2012d).
58. The total number of the most popular English words inside a specific Web page. We based on the popular English words that mentioned in (Wahsheh, *et al*, 2012d).

Figure 4.6 presents an algorithm of the enhanced Arabic content-based Web spam analyzer.

Algorithm	**Enhanced Arabic content-based Web spam analyzer.**
Input:	**List of URLs stored on a text file (ContentbasedURL.txt).**
Output:	**Table of the number of Web page features, stored in the database of the Arabic Web spam detection system.**

BEGIN

WHILE NOT EOF (ContentbasedURL.txt)

Read the URL of Web page.

Download a Web page.

Count the number of meaningless English/Arabic words in the <body> element.

Count the total number of the characters in the all <Meta> elements.

Count the number of the characters in the each <Meta> elements.

Count the number of Arabic/English words in each <Meta> elements.

Count the total number of Arabic words in all <Meta> elements.

Count the total number of English words in all <Meta> elements.

Count the total number of Symbol words in all <Meta> elements.

Count the total number of Symbol characters in all <Meta> elements.

Count the total number of characters of the symbols in the <body> element.

Count the total number of Arabic characters in a specific Web page.

Count the total number of Arabic characters in the <body> element.

Count the total number of characters of the English in a specific Web page.

Count the total number of English characters in the <body> element.

Count the total number of words with symbols in a specific Web page.

Count the minimum length of English word inside <body> element.

Count the minimum length of Arabic word inside <body> element.

Count the minimum length of English word inside <body> element.

Count the minimum length of symbol word inside <body> element.

Count the maximum length of Arabic/English word inside <body> element.

Count the maximum length of Arabic word inside <body> element.

Count the maximum length of English word inside <body> element.

Count the maximum length of symbol word inside <body> element.

Count the maximum length of symbol word inside <page>.

Compute the average lengths of words inside <body> element.

Compute the average lengths of Arabic words inside <body> element.

Compute the average lengths of English words inside <body> element.

Compute the average lengths of symbol words inside <body> element.

Count the total number of <Meta> elements inside a specific Web page.

Count Web page size (Kilo bytes).

Count the number of characters with the URL.

Count the complexity factor in <body> element.

Count the complexity factor in a specific Web page.

Count the total number of Arabic/English words inside <body> element.

Count the total number of Arabic/English words inside <title> element.

Count the total number of Arabic words inside <body> element.

Count the total number of Arabic words inside <title> element.

Count the total number of English words inside <body> element.

Count the total number of English words inside <title> element.

Count the total number of symbol words inside <body> element.

Count the total number of symbol words inside <title> element.

Count the visible page fraction inside a specific Web page.

Count the visible page fraction inside the <body> element.

Count Web page hidden text size (Kilo bytes) inside a specific Web page.

Count the total size of compressed files inside a specific Web page.

Count the total size of compression ratio inside a specific Web page.

Count the total number of unique Arabic/English words inside <body> element.

Count the total number of unique Arabic/English words inside a specific Web page.

Count the total number of unique Arabic words inside <body> element.

Count the total number of unique Arabic words inside a specific Web page.

Count the total number of unique English words inside <body> element.

Count the total number of unique English words inside a specific Web page.

Count the total number of unique symbol words inside <body> element.

Count the total number of unique symbol words inside a specific Web page.

```
    Count the number of image/images inside a specific Web page.
    Count the number of image-link /images-links inside a specific Web page.
    Count the total number of popular Arabic words inside a specific Web page.
    Count the total number of popular English words inside a specific Web page.
  END WHILE
END
```

Figure 4.6: Enhanced Arabic content-based Web spam analyzer

It should be noted that there are some of content-based features computed two times, one will be for the page <page>, and another for the <body>. This means that we try to monitor all the parts of HTML elements, especially out side the <body>. This can provide the benefits when we discuss the cloaking sections.

4.3.2 Link-based features

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

- Spam link farm. The spam link farm as we mentioned in the literature create heavily connected Web pages, in order to mislead search engines through the manipulation in the number of internal links and external links in the Web page (Du, et al, 2007).

- Using the expired domains. Spammers take the benefits of expired domains by inserting their spammed Web pages in. Also the spammers try to trick the users, when they used the names which are similar to the popular trusted and reputable domains names (Gyongyi & Garcia-Molin, 2005).

- Link spam comments in the blogs: The spammers may post the links to spam Web pages as a comment to the blogs. So the spam comments will increase the traffic on the honey spam blogs and forums (Niu, *et al*, 2006).

Wahsheh et al. (2012c) exhibits that the spammers used two types of link-based techniques in Arabic Web pages. The first type is based on using the links spam farms, which manipulate internal and external links in the Web pages in order to increase the rank of these spam Web pages. While the second type is based on using the expired domains which try to trick the users, when the spammers used the names which are similar to the popular trusted and reputable domains.

Figure 4.7 shows an example of Arabic link-based spam Web page which has many anchor text points to the same Web page without adding any valuable information.

Figure 4.7: Example of Arabic link-based spam Web page

In this thesis we detected the two previous link-based types and other types through the proposed link-based features to detect the spam Web pages. So Our Web

page analyzer extracts the link-based features which mentioned in Wahsheh, *et al*, (2012c) beside those newly adopted features. The Web page analyzer extracts link-based features as shown below:

1. The number of external links within the page under consideration.
2. The number of internal links within the page under consideration.
3. The total number of links (the internal and external) within the page under consideration.
4. The URL length (total number of characters in URL. This feature can be considered as the same URL feature in the content, but it can be different when we have a redirected Web page).
5. The total number of broken links. It is also known as dead link within the page under consideration. The broken links are called broken, due they are no longer point to non spam Web pages, so they decrease the rank of a Web page (Martinez-Romo & Arauj, 2012).
6. The total number of redirected links within the page under consideration.
7. The total number of empty link text (links with out anchor text) within the page under consideration.
8. The total number of empty links (anchor text without links) within the page under consideration.

Figure 4.8 presents an algorithm of Arabic link-based Web spam analyzer.

Algorithm	**Arabic link-based Web spam analyzer.**
Input:	**List of URLs stored on a text file (linkbasedURL.txt).**
Output:	**Table of the number of Web page features, stored in the database of the Arabic Web spam detection system, with the appropriate label.**

BEGIN

WHILE NOT EOF (linkbasedURL.txt)

Read the URL of a Web page.

Download a Web page.

Count the number of external links in the download Web page.

Count the total number of internal links in the download Web page.

Count the number of all links (the internal and external) in the download Web page.

Measure the URL length (number of characters).

Count the total number of broken links in the download Web page.

Count the total number of redirected links in the download Web page.

Count the total number of empty link text in the download Web page.

Count the total number of empty links in the download Web page.

END WHILE

END

Figure 4.8: Developed Arabic link-based Web spam analyzer

Figure 4.9 presents another example of Arabic link-based spam Web page which is full of advertisements. This type of link-based Web spam is called scraper Web page that does not contain any real content related to the Website, where the links redirect users to other Web sites.

Figure 4.9: Example of scraper Arabic link spam Web page

4.3.3 Cloaking features

Cloaking spam Web pages is based on the basic idea of producing two different versions of each Web page. The difference between them affected by the factors of content-based, link-based features, and the quality of the Web page. The high quality version appearing on the Web crawler to get the rank that actually contrary to the quality version appearing in the user browser (Lin.2009).

The Web page analyzer computes the difference between the content-based and link-based features which appears in the user browser; and the content-based and link-based features which appears in the Web crawler. Figure 4.10 presents the algorithm of Arabic cloaking Web spam analyzer.

Algorithm	**Cloaking Web spam analyzer.**
Input:	**List of URLs stored on a text file (CloakingWebspam.txt).**
Output:	**Table of the number of Web page (content-based and link-based) features, stored in the database of the Arabic Web spam detection system, and the decision as a (spam/ non-spam).**

BEGIN

WHILE NOT EOF (CloakingURL.txt)

Read the URL of a Web page.

Download a Web page.

Request Web page analyzer (content and link features).

Compute link/content features in the user browser.

Compute link/content features in the Web crawler.

Compute the difference values of link/content features (user browser-Web crawler).

END WHILE

END

Figure 4.10: Developed Arabic cloaking Web spam analyzer

Figures 4.11 and 4.12 present two versions of the same Web page as an example of cloaking Arabic spammed Web page.

Figure 4.11: Example of user browser version of Arabic cloaking Web page

Figure 4.11 shows the content that appears in the user browser version, while Figure 4.12 presents the content appearing in the Web crawler for the same Web page.

Figure 4.12: Example of Web crawler version of Arabic cloaking Web page

4.3.4 Content/link features

In this section, the Web page analyzer extracts the content-based and link-based features, which mentioned in section 4.3.1, and the section 4.3.2 respectively. The Web page analyzer export the content/link features to the CSV file, or stored them in the database to be used by the Arabic content/link Web spam detection system.

4.4 Apply classification algorithms

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

4.4.1 Logistic Regression algorithm

Logistic Regression is one type of regression analysis types. It is widely used statistical modeling technique for predicting the outcome of categorizing the variables

depending on the predictor variables. It can be either binomial or multinomial regression. The binomial (binary) can observe the outcome with two possible types as a (0, or 1) which expressed the straightforward interpretation, while the multinomial regression indicate that the outcomes can have more than two possible types (Wang.2005).

4.4.2 K-Nearest Neighbour (K-NN) algorithm

K-Nearest Neighbor also known as IBK in Weka. It is considered as the simplest machine learning algorithms, and it is one of the lazy learning types. The classification decision based on the closest training objects values *K*, which starts from 1, and indicates to the space of the neighborhoods around the test pattern (Yang.2006)

4.4.3 Decision Tree algorithm

The Decision Tree is one of the common classification techniques available in WEKA. It is visualizes as a tree-like model or a graph of start decision on the root node to the leaves nodes. The decision taken by comparing the features values against some constants through different nodes paths (Xhemali, *et al*, 2009).

Decision Tree high speed and powerful way to express the tree structure. It is widely used in the studies to help identifying the strategy decision for specific goals (Witten & Frank, 2005).

CHAPTER FIVE

IMPLEMENTATIONS & EXPERIMENTAL RESULTS

In this chapter we extracted the features of Arabic Web spam dataset which are mentioned in section 4.2, we used three groups within one dataset with various percentages of spam Web pages (2%, 30%, and 40%).

We apply three classifiers (Logistic Regression, Decision Tree, and *K-NN*) on these three groups of the spam dataset. Afterward we compare the results to identify the best classifier, capable to detect Arabic content/link Web spam. Finally extract the rules of the best classifier to build the Arabic content/link Web spam detection system.

5.1 Arabic content/link Web spam features extraction

This section is divided into three subsections. In the first subsection we extracted the content-based features and studied the spammer's behavior with the spam and non spam Web pages. In the second subsection we extracted the link based features. Finally in the third subsection we extracted the cloaking features.

5.1.1 Content-based features extraction

We extract the content-based features using our Web page Analyzer, and we have found clearly the spammers behavior within the set of features, which will be effective and have the main role in Arabic Web spam detection.

Figure 5.1 shows the spammer behaviors with the title element in Arabic Web pages compared with the non spam Web pages.

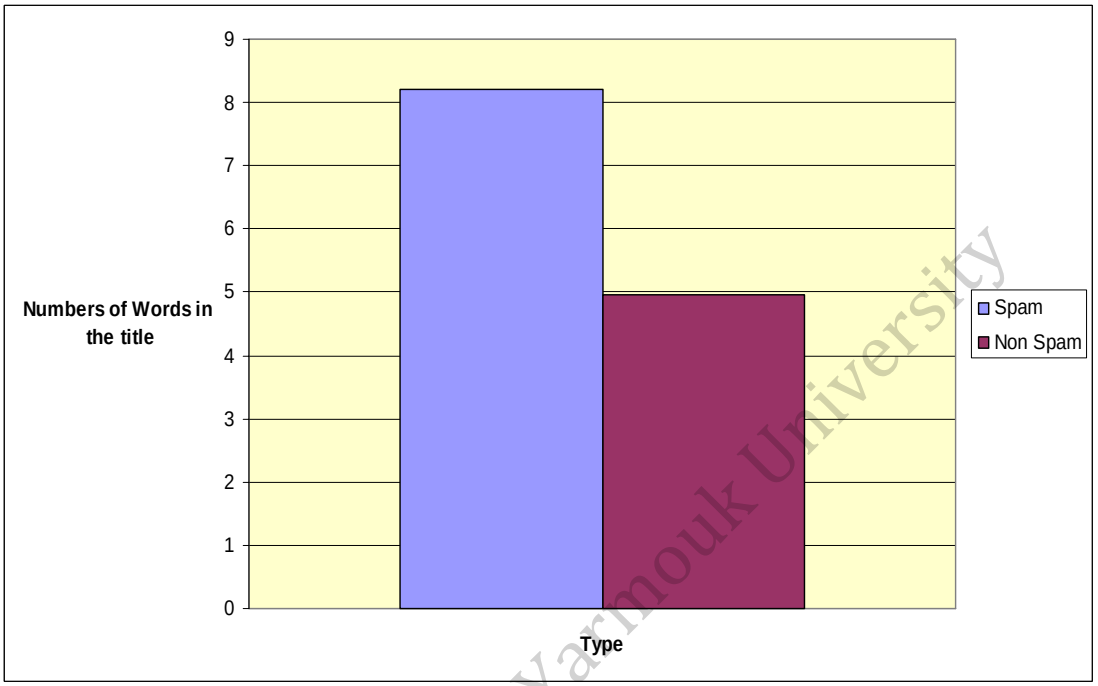


Figure 5.1: Spam behavior using title element

It is known that the increasing number of words in the <title> will lead the Web page to gain a heavier weight than it's really deserve on different search engines. Figure 5.1 shows the spammers behavior using the (Arabic, English, and symbols) words in the <title> element. The spam behavior stuff many keywords in the <title> element to increase the visibility in SERP. The threshold used is up to three duplications, if it exceeds three, there is a downturn in terms of visibility (Wahsheh, *et al*, 2012a).

High number of images is another spam behavior discovered in spammed Arabic Web pages to increase the rank of the spam Web page within the SERP. When the images are used as hyperlinks, the spammers trick the users with the content of images, while when users click on the images they get a spam Web pages which contradict to the desired Web page. Figure 5.2 presents that the spammers use the images as hyperlinks frequently, as and more than the non spam Web pages.

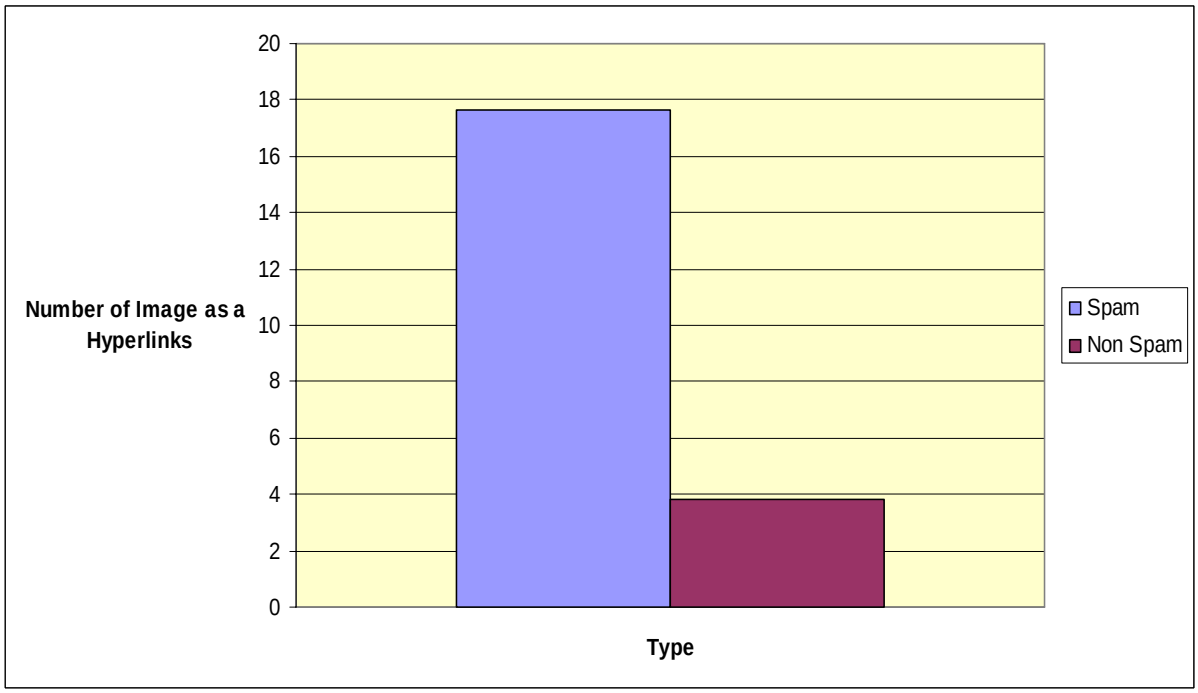


Figure 5.2: Spam behavior using images as a hyperlink

The spammers benefit from this spam behavior which considered as one of the pay per click (PPC) marketing techniques, by attracting more users to click on these spam images. So the spammers gain more revenues.

Figure 5.3 shows another influence content-based feature, where the spammers try to increase the length of Arabic words to increase the rank of spammed Web pages.

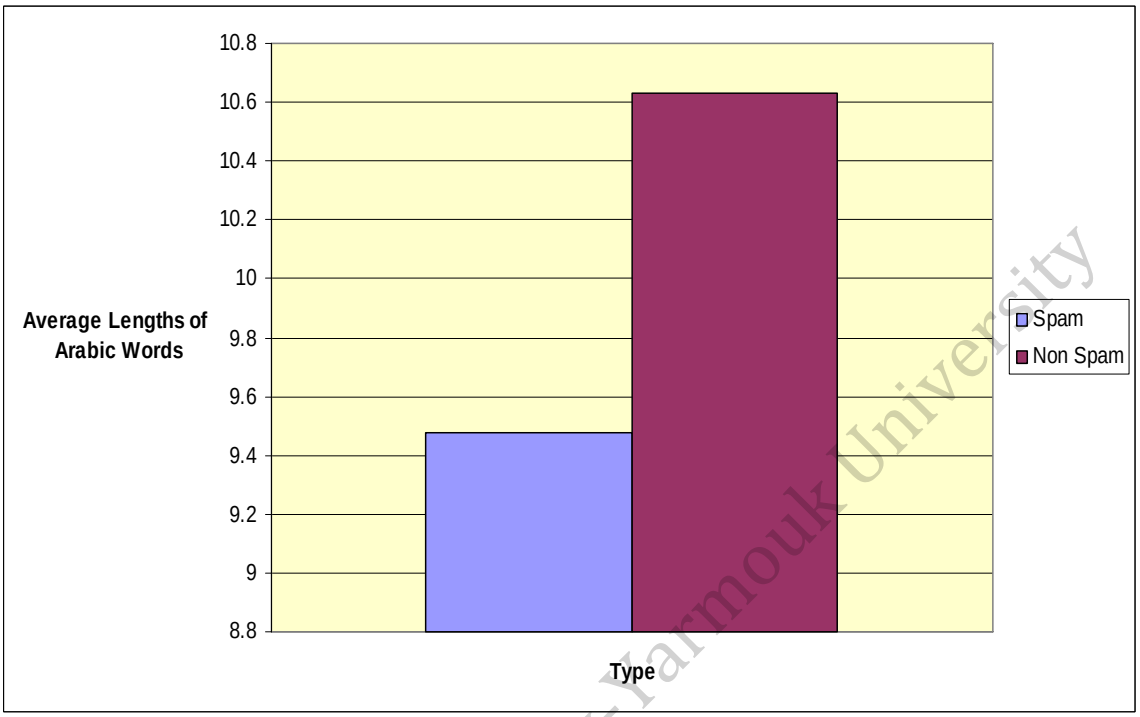


Figure 5.3: Spam behavior using Average lengths of Arabic words

5.1.2 Link-based features extraction

We extract the link-based features using our Web page analyzer. We have found difference of the spam factors using these features.

Figure 5.4 shows the distribution of the link-based features depending on the spam behavior.

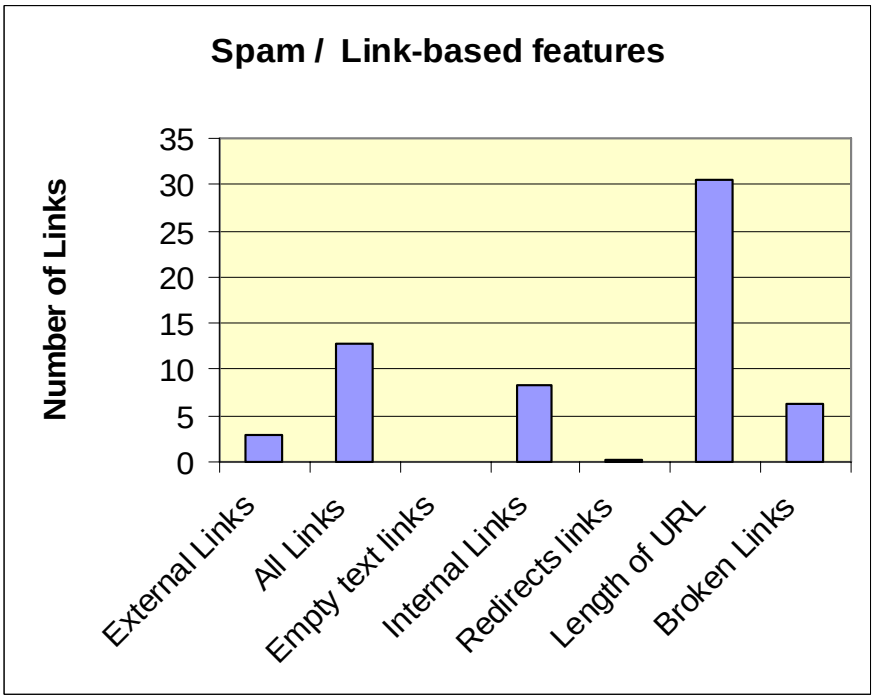


Figure 5.4: Spam behavior using link-based features

While Figure 5.5 shows the distribution of the link-based features depending on the non spam behavior.

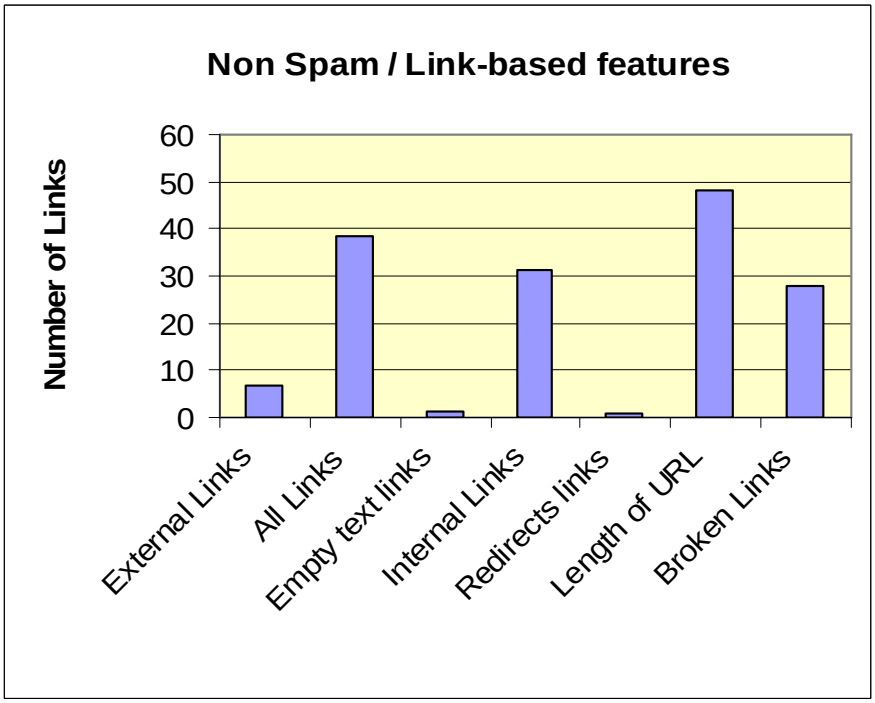


Figure 5.5: Non spam behavior using link-based features

5.1.3 Cloaking features extraction

We also extract the cloaking features using our Web page analyzer. We should note that if the difference result of the cloaking features is equal to zero, this means that we have differences between the user browser version and the Web crawler version. So the Web page under consideration is a spam Web page.

If we found that the cloaking features have negative numerical results, this means that the Web crawler version has more content and links than the user browser version. So this indicates that we have a spam behavior, with some exceptions for external links and broken links. We need first to identify the thresholds that determine if the difference between the two versions (the user browser version and the Web crawler version) is significant to identify a spam behavior or not. The thresholds depend on the contents/links of HTML elements, so if we have a duplicate in the content and high number of additional links, this means that we reach the threshold. While if the difference in the content/link features between the user browser version and the Web crawler version does not has any duplication, or additional links. This means that we have not reached the thresholds of a spam behavior.

High number of broken links and external links in the Web crawler version means that the Web page will not get a high rank. This contradicts to the spam behavior which tries to increase the rank of Web pages. So if we found that we have negative numerical results with the broken links and external links in the cloaking features, this means that we have non spam behavior. This may considered as an errors in designing the Web pages by Web masters.

If the cloaking features have positive numerical results; this means that the user browser version has more content/links than the Web crawler version. So we have a non

spam behavior, with some exceptions in external link and broken links. Increasing number of broken links and external links in the user browser version means that the version of the Web page that sent to the Web crawler does not contain the broken and external links. So these Web pages get a higher rank than it really deserve. Thus, the users will suffer from these links.

5.2 Apply the classifiers

In order to find the best Arabic content/link classification algorithms. Three classifiers (given in Weka) on the content/link Web spam dataset were tested.

We have several evaluation metrics which validates classification algorithms, such as:

1. Accuracy: Represents a fraction of training documents that assigned to the correct class by the classifier (Baeza-Yates & Ribeiro-Neto, 2010).

2. Error: Represents a fraction of training documents that assigned to the incorrect class by the classifier (Baeza-Yates & Ribeiro-Neto, 2010).

3. Precision (*P*): Represents a fraction of dividing the number of relevant retrieved documents over the total number of retrieved documents (Witten & Frank, 2005).

4. Recall (*R*): Represents a fraction of dividing the number of relevant retrieved documents over the total number of relevant documents (Witten & Frank, 2005).

5. True Positive (*TP*): Represents the number of items correctly labeled as belonging to the positive class (Witten & Frank, 2005).

6. False Positive (*FP*): Represents the number of items correctly labeled as belonging to the negative class (Witten & Frank, 2005).

7. *F-measure*: Is an accuracy measure that combine both the precision and recall values. The traditional *F-measure* formula is:

$$F - Measure = \frac{2PR}{P+R} \quad \text{.....(5.1)}$$

Where *P* is the Precision; *R* is the recall (Baeza-Yates & Ribeiro-Neto, 2010).

8. Receiver Operating Characteristic (ROC), or ROC curve: This curve depicts the performance of a binary classifier. It is plotting the fraction of True Positives Rate vs. the fraction of the False Positives Rate.

5.2.1 Content-based classification results

Applying the Logistic Regression algorithm on the three spam percentage groups yields accuracies of 98.0411%, 82.3529%, and 95.8333% respectively. The results of the accuracies and errors are shown in the Table 5.1.

Table 5.1. Content-based Logistic Regression Results

Spam percentage group	Accuracy	Error
2% spam group.	98.0411%	1.9589%
30% spam group.	82.3529%	17.6471%
40% spam group.	95.8333%	4.1667%

The results shown in Table 5.1 indicate that different percentages of spam with the three different dataset groups have a significant impact on the accuracy of the Logistic Regression classifier results.

Afterward we applied the second classifier; *K-NN* on the three spam percentage groups yields accuracies of 99.7293%, 85.2941%, and 95.8333% respectively. The results of the accuracies and errors are shown in the Table 5.2.

Table 5.2. Content-based *K-NN* (IBK) results (*K*=1)

Spam percentage group	Accuracy	Error
2% spam group.	99.7083	0.2917%
30% spam group.	85.2941%	14.7059%
40% spam group.	95.8333%	4.1667%

The results shown in Table 5.2 indicate that different percentages of spam with the three different dataset groups have a significant impact on the accuracy of the *K-NN* classifier results.

Finally we applied the Decision Tree classifier on the three spam percentage groups yield accuracies of 99.7611%, 88.1188%, and 96.875% respectively. The results of the accuracies and errors are shown in the Table 5.3.

Table5.3. Content-based Decision Tree Results

Spam percentage group	Accuracy	Error
2% spam group.	99.7611%	0.2389%
30% spam group.	88.1188%	11.8812%
40% spam group.	96.875%	3.125%

Tables 5.2 and 5.3 show clearly that the spam dataset group with the spam percentage of 2% achieved the highest results in detecting Arabic content-based Web spam, and the best classifier is Decision Tree. While we found that the spam dataset

group with the spam percentage of 30% achieved the lowest results in detecting Arabic content-based Web spam.

Table 5.4 shows accuracy comparisons between three groups of spam percentages of the dataset (2%, 30%, and 40%), with six previous Arabic content-based studies.

Table 5.4. Comparison of the accuracy values for content-based with six previous Arabic content-based studies

Classifier within group of spam percentages	True Positive Rate	False Positive Rate	Precision	Recall	F-measure	ROC
Decision Tree (2%) in this study.	0.998	0.12	0.998	0.998	0.998	0.987
Decision Tree (30%) in this study.	0.881	0.139	0.884	0.881	0.882	0.875
Decision Tree (40%) in this study.	0.969	0.013	0.972	0.969	0.969	0.983
K-NN (IBK) *K*=1 (2%) in this study.	0.997	0.136	0.997	0.997	0.997	1.0
K-NN (IBK) *K*=1 (30%) in this study.	0.853	0.032	0.92	0.853	0.867	0.911
K-NN (IBK) *K*=1 (40%) in this study.	0.958	0.017	0.964	0.958	0.959	0.971
Logistic Regression (2%) in this study.	0.98	0.98	0.961	0.98	0.971	0.846
Logistic Regression (30%) in this study.	0.824	0.038	0.912	0.824	0.842	0.935
Logistic Regression (40%) in this study.	0.958	0.017	0.964	0.958	0.959	0.973
K-NN K=1 (Wahsheh and Al-Kabi, 2011).	0.98	0.02	0.98	0.98	0.98	0.98
Decision Tree (Jarmah, *et al,* 2011).	0.91	0.01	-	-	-	-
Decision Tree (Al-Kabi, *et al,* 2011).	0.99	0.007	0.99	0.99	0.99	0.99
Decision Tree (Wahsheh, *et al,* 2012b).	0.98	0.003	0.99	0.99	0.99	0.99

Decision tree (15%) (Al-Kabi, *et al,* 2012).	1	0	1	1	1	1
Decision Tree (Wahsheh, *et al*. 2012a).	0.99	0.009	0.99	0.99	0.99	0.99

Table 5.4 shows clearly the superiority of Decision Tree algorithm in general, and especially within 2% spam percentage group, since it covers all content-based features of spammed Web pages.

5.2.2 Link-based classification results

The Logistic Regression also used in this study to classify spam link-based dataset. Using this classifier on the three spam percentage groups yields accuracies of 84.7924%, 79.3644%, and 76.5988% respectively. Table 5.5 shows the results of the accuracies and errors.

Table 5.5. Link based Logistic Regression results

Spam percentage group	Accuracy	Error
2% spam group.	84.7924%	15.2076%
30% spam group.	79.3644%	20.6356%
40% spam group.	76.5988%	23.4012%

The results shown in Table 5.5 indicates that Logistic Regression really fails to yield good accuracy results relative to the three spam percentage groups.

Applying the second classifier (*K-NN*) on the three spam percentage groups yields accuracies of 98.7174%, 99.7008%, and 95.8333%respectively. The results of the accuracies and errors are shown in the Table 5.6.

Table 5.6. Link-based *K-NN* (IBK) results (*K*=1)

Spam percentage group	Accuracy	Error
2% spam group.	98.7174%	1.2826%
30% spam group.	99.7008%	0.2992%
40% spam group.	98.7664%	1.2336%

The results shown in Table 5.6 indicate that the different percentages of spam with the three different dataset groups yields a high accuracy percentages to detect the spam link-based.

Finally we applied the Decision Tree classifier on the three spam percentage groups yields accuracies of 99.8174%, 99.7041%, and 99.6647% respectively. Table 5.7 presents the detailed results of the accuracies and errors.

Table 5.7. Link-based Decision Tree results

Spam percentage group	Accuracy	Error
2% spam group.	99.8174%	0.1826%
30% spam group.	99.7041%	0.2959%
40% spam group.	99.6647%	0.3353%

Table 5.6 and 5.7 show that the spam dataset group with the spam percentage of 2% achieved the highest results in detecting Arabic content-based Web spam, and the best classifier is Decision Tree.

Table 5.8 shows accuracy comparisons between the three percentages groups of spam (2%, 30%, and 40%), and previous Arabic link-based study.

Table 5.8. Comparison of the accuracy values with previous Arabic link-based study

Dataset	True Positive Rate	False Positive Rate	Precision	Recall	F-measure	ROC
Decision Tree (2%) in this study.	0.998	0.002	0.998	0.998	0.998	1.0
Decision Tree (30%) in this study.	0.997	0.003	0.997	0.997	0.997	1.0
Decision Tree (40%) in this study.	0.997	0.003	0.997	0.997	0.997	1.0
K-NN K=1 (2%) in this study.	0.998	0.002	0.998	0.998	0.998	1.0
K-NN K=1 (30%) in this study.	0.998	0.002	0.998	0.998	0.998	1.0
K-NN K=1 (40%) in this study.	0.998	0.002	0.998	0.998	0.998	1.0
Logistic Regression (2%) in this study.	0.848	0.513	0.834	0.848	0.828	0.822
Logistic Regression (30%) in this study.	0.794	0.307	0.793	0.794	0.784	0.823
Logistic Regression (40%) in this study.	0.766	0.289	0.766	0.766	0.76	0.823
Decision Tree (Wahsheh, *et al*, 2012c).	0.91	0.08	0.92	0.91	0.91	0.98

Table 5.8 shows the superiority of the Decision Tree when used 2% spam percentages.

5.2.3 Cloaking classification results

The Logistic Regression is also applied on the three spam percentage groups to classify cloaking dataset. Applying it on the three spam percentage groups yield accuracies of 97.9933%, 87.3737%, and 86.1644% respectively. Table 10 shows the results of the accuracies and errors of using Logistic Regression within cloaking spammed Web pages.

Table 5.9. Cloaking Logistic Regression results

Spam percentage group	Accuracy	Error
2% spam group.	97.9933%	2.0067%
30% spam group.	87.3737%	12.6263%
40% spam group.	86.1644%	13.8356%

The results shown in Table 5.9 indicate that we have not good percentages to detect the spam cloaking, on the three different percentages of spam. Table 5.10 shows the accuracy and error results of applying *K-NN* on the cloaking dataset. Applying the *K-NN* on the three spam groups, yield accuracy results of 98.3437%, 96.1014%, and 89.4434% respectively.

Table 5.10. Cloaking *K-NN* (IBK) results (*K*=1)

Spam percentage group	Accuracy	Error
2% spam group.	98.3437%	1.6563%
30% spam group.	96.1014%	3.8986%
40% spam group.	89.4434%	10.5566%

The results shown in Table 5.10 indicate that the three percentage groups of spam dataset gain a high percentage to detect Arabic link-based spammed Web pages.

Finally we applied the Decision Tree classifier on the three different percentages of spam, and it yields accuracies of 99.8174%, 96.1014%, and 89.4434% respectively. Table 5.11 shows the results of accuracies and errors.

Table 5.11. CloakingDecision Tree results

Spam percentage group	Accuracy	Error
2% spam group.	99.8174%	0.1826%
30% spam group.	96.1014%	3.8986%
40% spam group.	89.4434%	10.5566%

Table 5.11 shows the superiority of the Decision Tree when used the spam percentage of 2%.

In our spam dataset, we were considering the Web pages as a cloaking type if the Web page has a cloaking behavior in any of its HTML elements. Although the results showed superiority in detecting cloaking with the Decision Tree, when used the spam group of percentage 2%. It is not appropriate to extract the rules from the Decision Tree, due we need to make a check of every HTML element if it uses cloaking behavior or not. So we need to use all the rules not only the rules of the Decision Tree.

5.2.4 Content/link classification results

The vast majority of Arabic spam Web pages used a content-based techniques only, while other Arabic spam Web pages used the link-based techniques only. However there are Arabic spammed Web pages used both content and link spam techniques, so this need to combine the content and link features to detect these spammed Web pages which adopts both content and link spam techniques.

Therefore we present this section in order to detect the rules of Arabic content/link based Web spam, to improve the Arabic spam detection techniques.

We also applied the Logistic Regression with the content/link dataset on the three different percentages of spam, and it yields accuracies of 75.5894%, 70.2168%,

and 67.3343% respectively. The results of accuracies and errors are shown in the Table 5.12.

Table 5.12. Content/link Logistic Regression results

Spam percentage group	Accuracy	Error
2% spam group.	75.5894%	24.4106%
30% spam group.	70.2168%	29.7832%
40% spam group.	67.3343%	32.6657%

The results shown in Table 5.12 indicate that we have bad results to detect the content/link spam Web pages, through the three different groups with different spam percentages. Then as shown in Table 5.13 we applied the second classifier; *K-NN*, which yields an accuracy of 85.9931%, 86.9716%, and 88.2702% respectively. Table 5.13 shows the results of accuracies and errors.

Table 5.13. Content/link *K-NN* (IBK) results (*K*=1)

Spam percentage group	Accuracy	Error
2% spam group.	85.9931%	14.0069%
30% spam group.	86.9716%	13.0284%
40% spam group.	88.2702%	11.7298%

The results shown in Table 5.13 indicate that three different groups with different spam percentages achieved good results to detect the spam content/link Web spam.

Finally we applied the Decision Tree classifier on the spam dataset with three different groups with different spam percentages, and yielded 98.2422%, 97.0891%, and 96.7218% respectively. Table 5.14 presents the results of accuracies and errors.

Table 5.14. Content/link Decision Tree results

Spam percentage group	Accuracy	Error
2% spam group.	98.2422%	1.7578%
30% spam group.	97.0891%	2.9109%
40% spam group.	96.7218%	3.2782%

Table 5.14 shows the superiority of Decision Tree on a 2% spam dataset group. While Table 5.15 presents accuracy comparisons between three percentage groups of spam dataset (2%, 30%, and 40%) and shows the superiority of the Decision Tree when applied on 2% spam dataset group.

Table 5.15. Comparison of the accuracy values for content/link

Dataset	True Positive Rate	False Positive Rate	Precision	Recall	F-measure	ROC
Decision Tree (2%) in this study.	0.98	0.18	0.98	0.98	0.98	0.89
Decision Tree (30%) in this study.	0.97	0.26	0.97	0.97	0.96	0.85
Decision Tree (40%) in this study.	0.96	0.27	0.96	0.96	0.96	0.84
K-NN K=1 (2%) in this study.	0.86	0.17	0.92	0.86	0.88	0.87
K-NN K=1 (30%) in this study.	0.87	0.42	0.91	0.87	0.88	0.83
K-NN K=1 (40%) in this study.	0.88	0.26	0.91	0.88	0.89	0.81
Logistic Regression (2%) in this study.	0.75	0.47	0.87	0.75	0.80	0.70
Logistic Regression (30%) in this study.	0.70	0.43	0.86	0.70	0.75	0.69
Logistic Regression (40%) in this study.	0.67	0.42	0.85	0.67	0.73	0.69

5.3 Rules extraction

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

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

5.3.1 The programming language

We used the Java programming language to build the Arabic content/link Web spam detection system, depending on the extracted rules of Decision Tree.

Java programming language provides many libraries, with high performance to analyze the HTML elements. Therefore we use it to include all the tasks needed to build our system.

Figure 5.6 shows about tab of our Arabic content/link Web spam detection system.

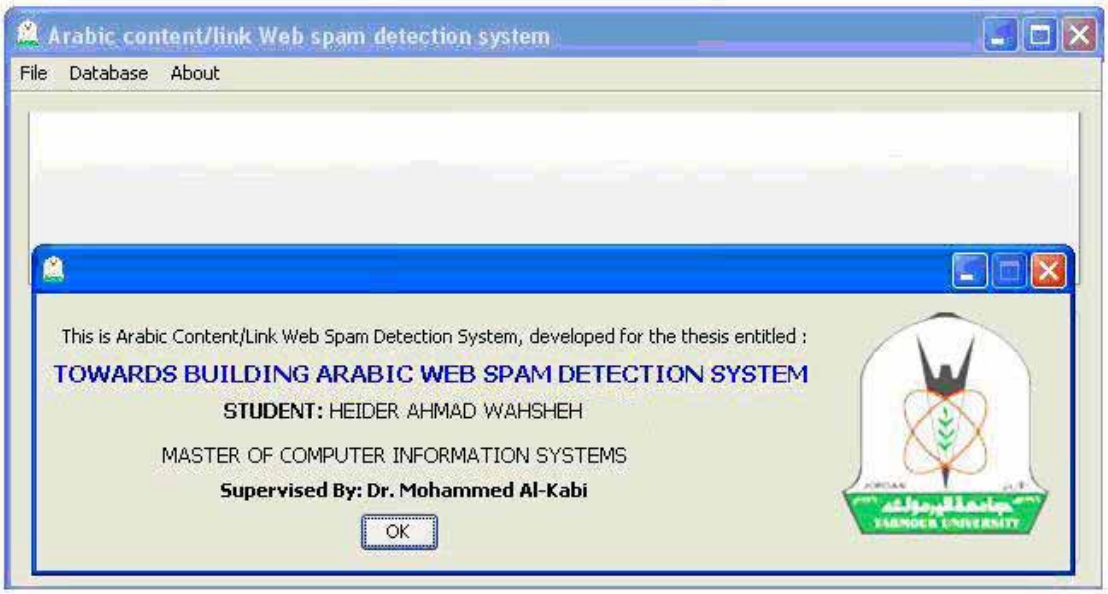


Figure 5.6: About interface of Arabic content/link Web spam detection system

5.3.2 Arabic content Web spam detection system

We found that the Arabic content-based Web spam can be detected by using the best content-based features which categorize it into six categories. The best categories compose the rules which were extracted from the Decision Tree when applied on (2%) spam percentage group.

Each category contains the following features:

1. The first category contains one content-based feature which checks if we have a number of meaningless (Arabic/English) in the HTML elements that can present the duplication or keywords stuffing techniques.

2. The second category check if we face a key stuffing technique, where the following two parameters are computed:

   2.1 Compute the difference between the total number of Arabic/English words inside the <body> element, and the total number of unique Arabic/English words inside the <body> element. If the results

greater than or equal two third of the total number of Arabic/English words inside <body>. This means that we have spam behavior.

2.2 Compute the difference between the total number of Arabic/English words inside a specific Web page, and the total number of unique Arabic/English words inside a specific Web page. If the results are greater than or equal two third of the total number of Arabic/English words inside a specific Web page. This means that we have spam behavior.

3. The third category contains a number of content-based features such as: the number of Arabic popular words, the size of compression ratio (in Kilo bytes), page size (in Kilo bytes), the maximum Arabic/English word length inside (<body>, or a specific Web page), the size of hidden text (in Kilo bytes), and the total number of images. Figure 5.7 presents the third category of content-based rule using Decision Tree on (2%) spam percentage group.

```
Number of Arabic popular words <= 22
|  Size of compression ratio <= 7.831862
|  |  page size (Kilo bytes) <= 11021
|  |  |  Maximum Arabic/English word length inside a specific Web page <= 27
|  |  |  |  Size of compression ratio <= 4.402923
|  |  |  |  |  Size of hidden text <= 72: spam
|  |  |  |  |  Size of hidden text > 72: non spam
|  |  |  |  Size of compression ratio > 4.402923: non spam
|  |  |  Maximum Arabic/English word length inside a specific Web page > 27: spam
|  |  page size (Kilo bytes) > 11,021: non spam
|  Size of compression ratio > 7.831862
|  |  Maximum Arabic/English word length inside a specific Web page <= 9: spam
```

```
| | Maximum Arabic/English word length inside a specific Web page > 9: non spam
Number of Arabic popular  > 22
| Number of images <= 159: spam
| Number of images > 159: non spam
```

Figure 5.7: Content-based rules of the third category using Decision Tree on (2%) spam percentage group.

4. The fourth category contains many content-based features as follows: the maximum Arabic word length inside (<body>, or a specific Web page), the average lengths of Arabic/English words inside the (<body>, or a specific Web page), the average lengths of Arabic words inside the (<body>, or a specific Web page), maximum Arabic/English word length inside a specific Web page, total number of characters of the symbols in all <Meta>, average length of English words inside a specific Web page, average length of Symbol words inside a specific Web page, number of unique Symbol words inside a specific Web page, and number of English popular words. Figure 5.8 presents the fourth category of the content-based rules which were extracted from Decision tree (2%) with many considerations.

```
Maximum Arabic word length inside a specific Web page <= 52
| Maximum Arabic/English word length inside <body> <= 5.401695
| | Average length of Arabic words inside the <body> <= 5.416667
| | | Average lengths of Arabic/English words inside the <body> <= 5.243781: non spam
| | | Average length of Arabic words inside the <body> > 5.243781
| | | | Maximum Arabic/English word length inside a specific Web page <= 20: spam
| | | | Maximum Arabic/English word length inside a specific Web page > 20: non spam
| | Average lengths of Arabic words inside a specific Web page > 5.416667: spam
| Average lengths of Arabic/English words inside the <body> > 5.401695
```

```
| | Number of characters of the symbols in all <Meta> <= 26
| | | Average length of English words inside a specific Web page <= 5.423879
| | | | Maximum Arabic/English word length inside a specific Web page <= 9: spam
| | | | Maximum Arabic/English word length inside a specific Web page > 9: non spam
| | | Average length of English words inside a specific Web page > 5.423879
| | | | Average length of Symbol words inside a specific Web page <= 25.346215: non spam
| | | | Average length of Symbol words inside a specific Web page <= 25.346215: non spam > 25.346215
| | | | | Number of unique Symbol words inside a specific Web page <= 140
| | | | | | Number of English popular words <= 2: non spam
| | | | | | Number of English popular words > 2: spam
| | | | | | Number of unique Symbol words inside a specific Web page > 140: non spam
| | Number of characters of the symbols in all <Meta> > 26
| | | | Number of unique Symbol words inside a specific Web page <= 12: spam
| | | | Number of unique Symbol words inside a specific Web page > 12
| | | | Number of English popular words <= 4
| | | | | Maximum Arabic/English word length inside a specific Web page <= 20: non spam
| | | | Maximum Arabic/English word length inside a specific Web page > 20: spam
| | | | Number of English popular words > 4: non spam
Maximum Arabic word length inside a specific Web page > 52: spam
```

Figure 5.8: Content-based rules of the fourth category using Decision Tree on (2%) spam percentage group.

5. The fifth category contains other influential content-based features as follows: number of images-links inside a specific Web page, maximum symbol word length inside a specific Web page, number of Arabic/English words inside <title>, compressed files inside a specific Web page, number of English characters inside <Meta>, number of English characters inside a specific Web page, number of characters of the symbols in all <Meta>, and the number of unique symbol words inside a specific Web page.

Figure 5.9 shows a fifth category of the content-based rules were extracted using Decision Tree on (2%) spam percentage group.

```
Number of images-links inside a specific Web page <= 13
|  Maximum Symbol word length inside a specific Web page <= 49
|  |  Number of Arabic/English words inside <title> <= 9
|  |  |  Number of images-links inside a specific Web page <= 3
|  |  |  |  Compressed files inside a specific Web page <= 1,771
|  |  |  |  |  Number of English characters inside <Meta> <= 28
|  |  |  |  |  |  Number of Arabic/English words inside <title> <= 4: spam
|  |  |  |  |  |  Number of Arabic/English words inside <title> > 4: non spam
|  |  |  |  |  Number of English characters inside <Meta> > 28: non spam
|  |  |  |  Compressed files inside a specific Web page > 1,771
|  |  |  |  |  Number of English characters inside a specific Web page <= 660
|  |  |  |  |  |  Number of symbol characters inside <Meta> <= 35
|  |  |  |  |  |  |  Number of Arabic/English words inside <title> <= 4: non spam
|  |  |  |  |  |  |  Number of Arabic/English words inside <title> > 4
|  |  |  |  |  |  |  |  Number of Arabic words inside <title> <= 4
|  |  |  |  |  |  |  |  |  Number of Arabic/English words inside <title> <= 6: spam
|  |  |  |  |  |  |  |  |  Number of Arabic/English words inside <title> > 6: non spam
|  |  |  |  |  |  |  |  Number of Arabic words inside <title>> 4: non spam
|  |  |  |  |  |  Number of symbol characters inside <Meta> > 35
|  |  |  |  |  |  |  Number of Arabic/English words inside <title> <= 5: spam
|  |  |  |  |  |  |  Number of Arabic/English words inside <title> > 5: non spam
|  |  |  |  |  Number of symbol characters inside <Meta> > 660
|  |  |  |  |  |  Number of symbol characters inside <Meta> <= 2006
|  |  |  |  |  |  |  Number of Arabic/English words inside <title>  <= 5: non spam
|  |  |  |  |  |  |  Number of Arabic/English words inside <title> > 5: spam
|  |  |  |  |  |  Number of English characters inside a specific Web page > 2006: non spam
|  |  |  Number of images-links inside a specific Web page > 3
|  |  |  |  Number of Arabic/English words inside <title>  <= 6: spam
|  |  |  |  Number of Arabic/English words inside <title> > 6: non spam
```

```
| | Number of Arabic/English words inside <title> > 9
| | | Number of English characters inside a specific Web page <= 21: non spam
| | | Number of English characters inside a specific Web page > 21: spam
| Maximum symbol word length inside a specific Web page > 49
| | Number of unique symbol words inside a specific Web page <= 10
| | | Maximum symbol word length inside a specific Web page <= 81: non spam
| | | Maximum symbol word length inside a specific Web page > 81
| | | | Number of Arabic/English words inside <title> <= 4: spam
| | | | Number of Arabic/English words inside <title> > 4: non spam
| | Number of unique symbol words inside a specific Web page > 10: non spam
Number of images-links inside a specific Web page > 13
| Number of unique symbol words inside a specific Web page <= 25: spam
| Number of unique symbol words inside a specific Web page > 25: non spam
```

Figure 5.9: Content-based rules of the fifth category using Decision Tree on (2%) spam percentage group

6. The sixth category contains the last influence content-based features as follows: number of <Meta> elements inside a specific Web page, number of characters within the URL, number of Arabic/English characters inside a specific Web page, average lengths of Arabic/English words inside a specific Web page, average length of Arabic words inside a specific Web page, number of <Meta> elements inside a specific Web page, number of Arabic words in each <Meta> elements, number of Arabic characters inside <Meta>, number of Arabic/English characters inside <Meta>, number of Arabic/English characters inside a specific Web page.

Figure 5.10 presents the sixth category of the content-based rules using Decision Tree on (2%) spam percentage group.

Number of <Meta> inside a specific Web page <= 10

| Number of characters within the URL <= 20

| | Number of characters within the URL <= 18

| | | | Number of Arabic/English characters inside a specific Web page <= 14

| | | | Number of characters within the URL <= 17

| | | | | Average length of Arabic/English words inside a specific Web page <= 8.101795: non spam

| | | | | Average length of Arabic/English words inside a specific Web page > 8.101795: spam

| | | | Number of characters within the URL > 17

| | | | | Average length of Arabic words inside a specific Web page <= 6.040346

| | | | | | Average length of Arabic/English words inside a specific Web page <= 14.197389

| | | | | | | Number of <Meta> inside a specific Web page <= 2: non spam

| | | | | | | Number of <Meta> inside a specific Web page > 2

| | | | | | | | Average length of Arabic/English words inside a specific Web page <= 6.630199

| | | | | | | | | Number of Arabic characters in each <Meta> <= 0: spam

| | | | | | | | | Number of Arabic characters inside <Meta>> 0: non spam

| | | | | | | | Average length of Arabic/English words inside a specific Web page > 6.630199

| | | | | | | | | Average length of Arabic words inside a specific Web page <= 5.863636: non spam

| | | | | | | | | Average length of Arabic words inside a specific Web page > 5.863636

| | | | | | | | | | Average length of Arabic/English words inside a specific Web page <= 7.586346: non spam

| | | | | | | | | Average length of Arabic/English words inside a specific Web page > 7.586346

| | | | | | | | | | Number of characters with the URL <= 89: spam

| | | | | | | | | | Number of characters with the URL > 89: non spam

| | | | | | Average length of Arabic/English words inside a specific Web page > 14.197389

| | | | | | | Number of Arabic/English characters inside <Meta><= 57: spam

| | | | | | | Number of Arabic/English characters inside <Meta> > 57: non spam

| | | | | Average length of Arabic words inside a specific Web page > 6.040346: non spam

| | | Number of Arabic/English characters inside a specific Web page > 14

| | | | Number of characters with the URL <= 65: spam

| | | | Number of characters with the URL > 65: non spam

| | Number of characters with the URL > 18: spam

| Number of characters with the URL > 20: non spam

```
Number of <Meta> inside a specific Web page > 10
|  Number of Arabic/English characters inside <Meta><= 26: spam
|  Number of Arabic/English characters inside <Meta>> 26: non spam
```

Figure 5.10: Content-based rules of the sixth category using Decision Tree on (2%) spam percentage group

The Figure 5.11 shows the algorithm of Arabic content-based Web spam detection. The algorithm used the six content-based rules (mentioned in 5.3.2)

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

```
BEGIN
  Open ContentbasedURl.txt or Database
  While Not EOF (ContentbasedURl.txt)
    Read the URL of Web page
    Download a Web page.
    Call content-based Web spam detection algorithm.
    Call First category of the Content-based Rules.
    Call second category of the Content-based Rules.
    Call third category of the Content-based Rules.
    Call fourth category of the Content-based Rules.
    Call fifth category of the Content-based Rules.
    Call sixth category of the Content-based Rules.
     Make a decision of non spam/ or true percentage of spam.
  END WHILE
END
```

Figure 5.11: Arabic content-based Web spam detection system

5.3.3 *Arabic link Web spam detection system*

Depending on the rules which were extracted from the Decision Tree when applied on (2%) spam percentage group, we found that the Arabic link-based Web spam can be detected by using best link-based features (not all features which were extracted through link-based Web analyzer). The following link-based features considered as a best link-based features: number of characters within the URL, number of empty link text, number of external links, number of broken links, number of internal links, number of redirected links.

Figure 5.12 presents the link-based rules using Decision Tree applied on (2%) spam percentage group.

```
Number of characters within the URL <= 28
|  Number of empty link text <= 0
|  |  Number of internal links <= 17
|  |  |  Number of external links <= 1
|  |  |  |  Number of broken links <= 11
|  |  |  |  |  Number of internal links <= 9
|  |  |  |  |  |  Number of internal links <= 3
|  |  |  |  |  |  |  Number of internal links <= 0: non spam
|  |  |  |  |  |  |  Number of internal links > 0: spam
|  |  |  |  |  |  Number of internal links > 3: non spam
|  |  |  |  |  Number of internal links > 9: spam
|  |  |  |  Number of broken links > 11: non spam
|  |  |  Number of external links > 1: spam
|  |  Number of internal links > 17: non spam
|  Number of empty link text > 0: non spam
Number of characters within the URL > 28
|  Number of internal links <= 1
|  |  Number of redirected links <= 0
```

| | | Number of internal links <= 0: non spam

| | | Number of internal links > 0

| | | | Number of characters within the URL <= 41: non spam

| | | | Number of characters within the URL > 41: spam

| | Number of redirected links > 0: spam

| Number of internal links > 1: non spam

Figure 5.12: Link-based rules using Decision Tree applied on (2%) spam percentage group

Figure 5.13 presents the algorithm of Arabic link-based Web spam detection.

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

BEGIN

Open LinkbasedURl.txt or Database

While Not EOF (LinkbasedURl.txt)

Read the URL of Web page

Download a Web page.

Call link-based Web spam detection algorithm.

Call Link-based Rules.

Make a decision of non spam/ or true percentage of spam.

END WHILE

END

Figure 5.13: Arabic link-based Web spam detection system

5.3.4 Arabic content/link Web spam detection system

Using the rules which extracted from the Decision Tree (2%) spam percentage, for the content-based and link-based, we merge the two algorithms and built the Arabic content/link Web spam detection system.

The Figure 5.14 shows the algorithm of Arabic content/link-based Web spam detection system.

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

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

Figure 5.14: Arabic content/link-based Web spam detection system

5.3.5 Arabic cloaking Web spam detection system

The results of the classified cloaked Web pages recommended that it is necessary to check every cloaking feature in the Web page, when we want to detect the cloaking behavior.

To extract all the cloaking features we need to extract all content/link features then find the difference between the two copies (user browser and Web crawler) of the cloaked Web page.

As we mentioned in section 5.1.3 (Cloaking features extraction), the cloaking features numerical values determine if we have a spam behavior or not.

The algorithm of Arabic cloaking Web spam detection system is too long, so we summarize the main steps of it, as shown in Figure 5.15.

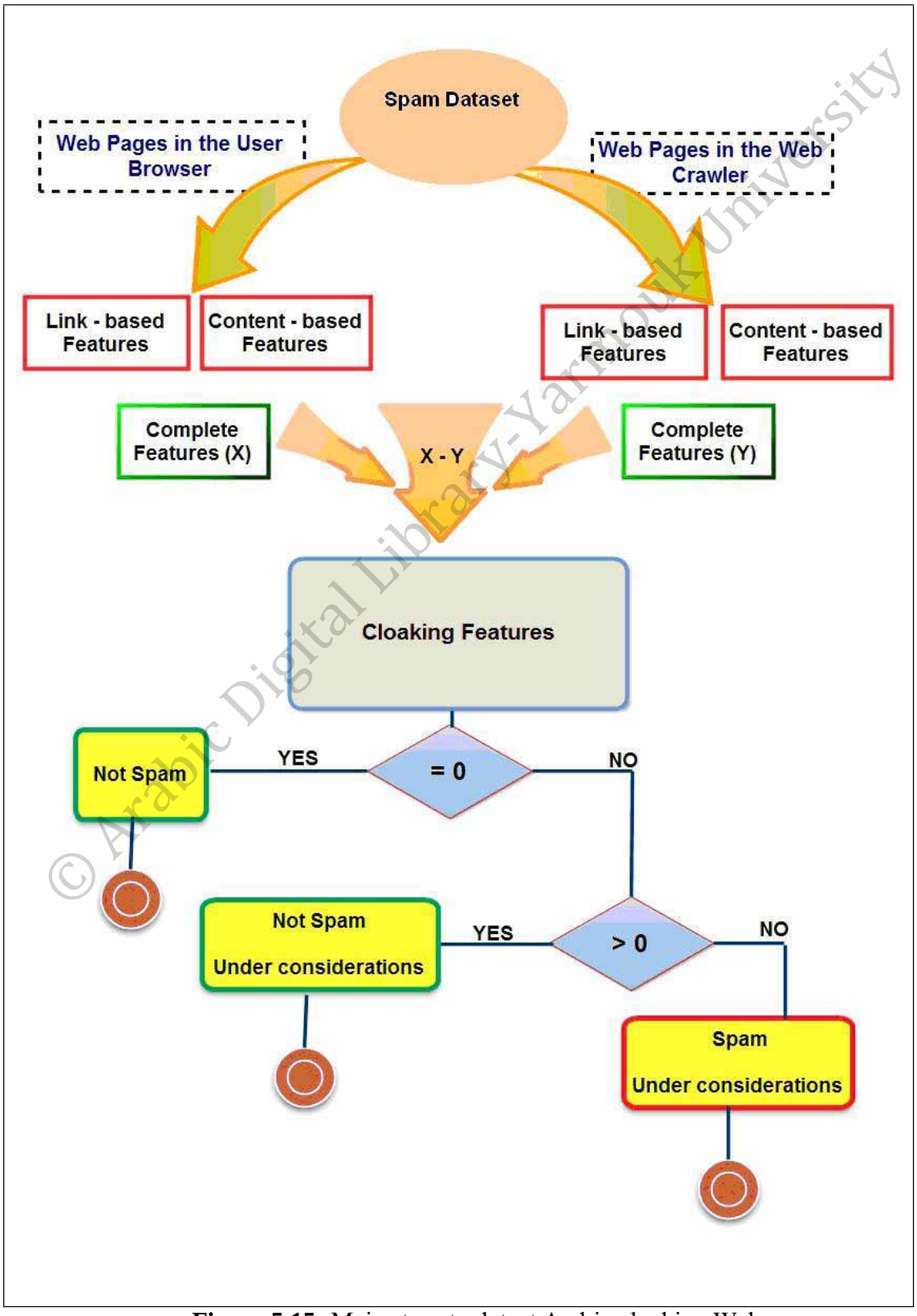


Figure 5.15: Main steps to detect Arabic cloaking Web spam

CHAPTER SIX

EVALUATION RESULTS

In this chapter we will evaluate all types of our Arabic content/link Web spam detection system, by using Decision Tree classifier.

We used 23,000 Arabic Web pages of our Web spam dataset as training dataset. The rules of the best classifier (the Decision tree) when the spam percentage 2% were extracted. Beside our training spam dataset we used around 5,000 Arabic spam and non spam Web pages as a test dataset. The test dataset labeled as spam and non spam. We use it to evaluate the effectiveness and performance of our Arabic content/link Web spam detection system. Then we used the Weka for the second time (the first time was in section 5.2 Apply the classifiers), and applied the Decision Tree classifier to evaluate our Arabic content/link Web spam detection system. Figure 6.1 shows the main menu of Arabic content/link Web spam detection system.

Figure 6.1: Main menu of Arabic content/link Web spam detection system

6.1 Evaluating Arabic content-based Web spam detection system

We used 5,000 spam and non spam Web pages as test dataset to evaluate the Arabic content-based Web spam detection system. The results of testing our Arabic content-based Web spam detection system yields an accuracy of 90.1099% in detecting content-based Web spam. Figure 6.2 shows the running of Arabic content-based Web spam detection system.

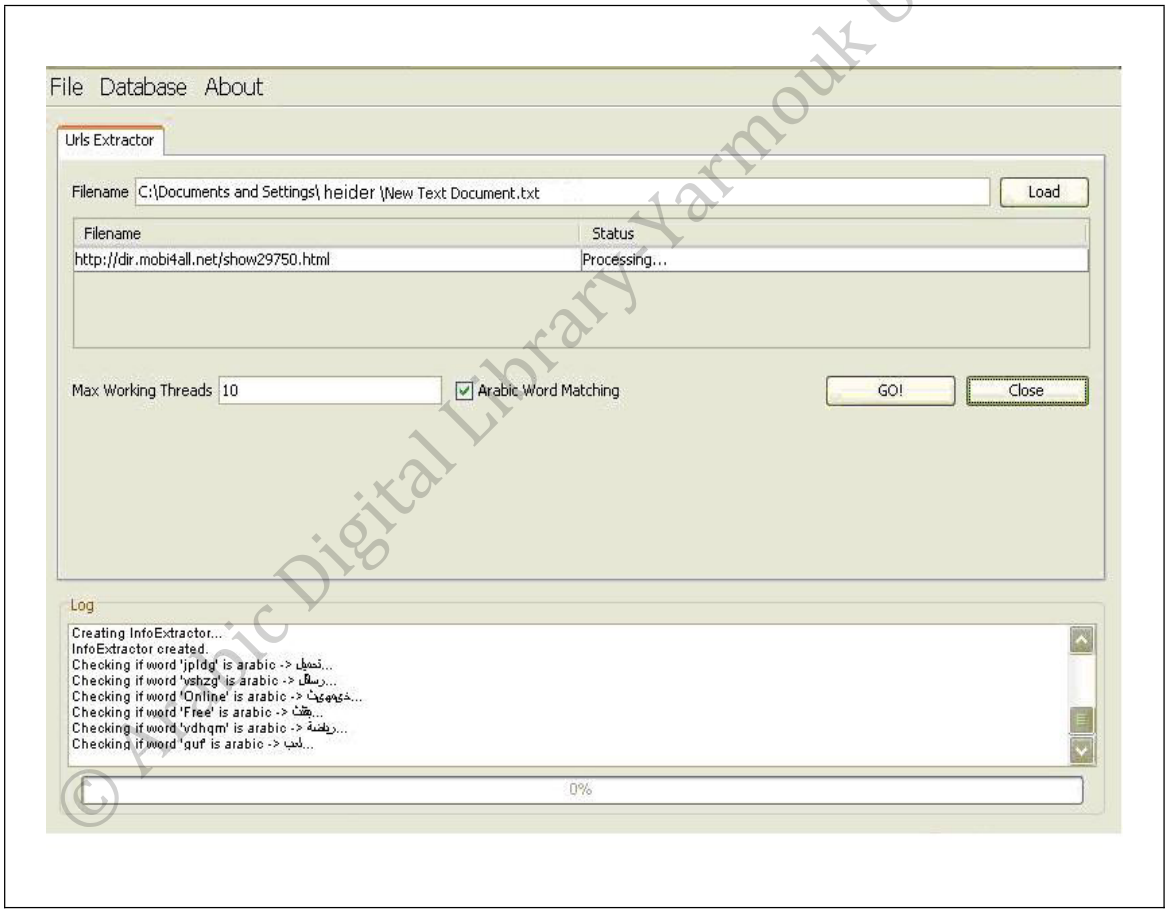


Figure 6.2: Running Arabic content-based Web spam detection system

Figure 6.3 presents an evaluation process of our Arabic content-based web spam detection system.

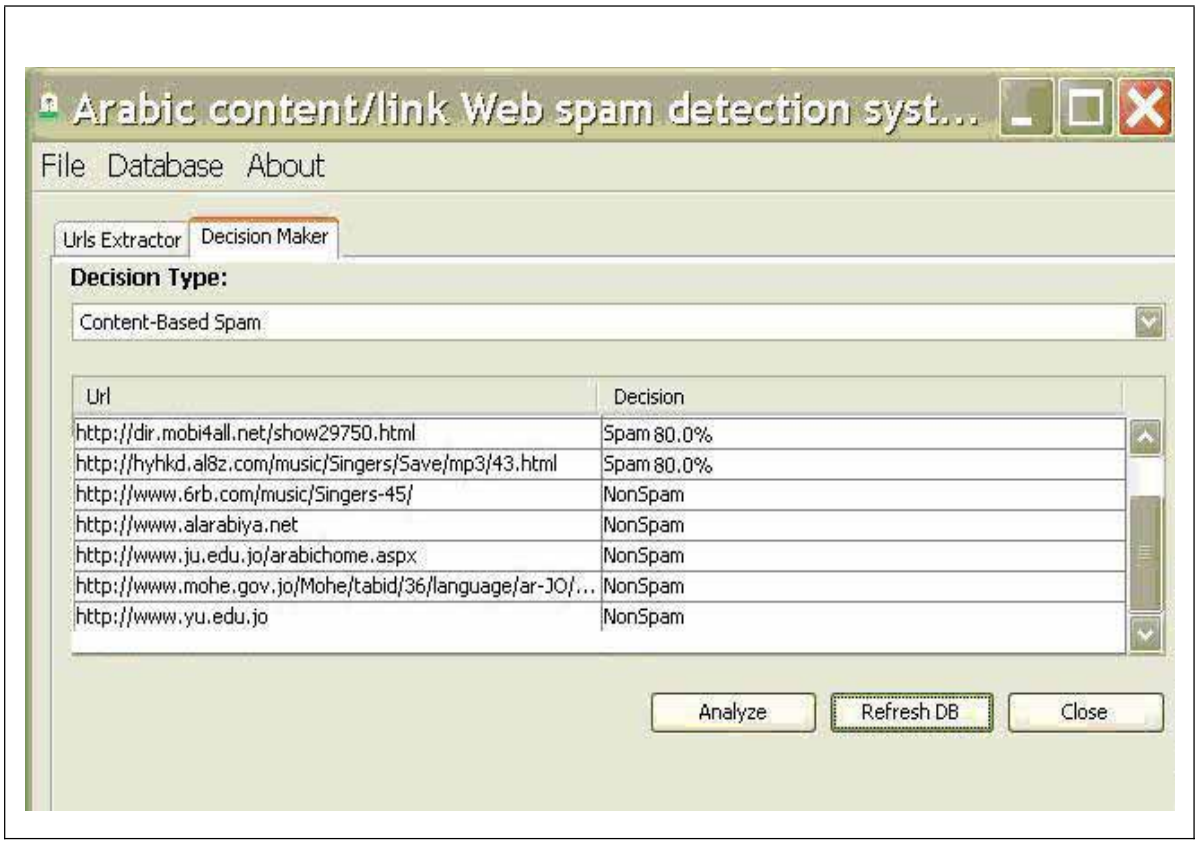


Figure 6.3: Evaluation process of Arabic content-based Web spam detection system

Table 6.1 presents the detailed evaluation results' of our Arabic content-based Web spam detection system, with accuracy, error, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure, and Receiver Operating Characteristic (ROC).

Table 6.1. Evaluation results of Arabic content-based Web spam detection system

Test Dataset	Accuracy	Error	True Positive Rate	False Positive Rate	Precision	Recall	F-Measure	ROC
Spam	-	-	0.974	0.525	0.915	0.974	0.944	0.882
Non spam	-	-	0.475	0.026	0.76	0.475	0.585	0.882
All	90.1099%	9.8901%	-	-	-	-	-	-

6.2 Evaluating Arabic link Web spam dtection system

We used the same test dataset which consists of 5,000 Arabic Web pages to evaluate the Arabic link-based Web spam detection system. Arabic link-based Web spam detection system yields 93.1034% accuracy in detecting link-based Web spam. Figure 6.4 shows the evaluations process of our Arabic linked-based Web spam detection system.

Figure 6.4: Evaluation process of our Arabic link-based Web spam detection system

Table 6.2 shows the detailed evaluation results' of our Arabic link-based Web spam detection system.

Table 6.2. Evaluation results of Arabic link-based Web spam detection system

Test Dataset	Accuracy	Error	True Positive Rate	False Positive Rate	Precision	Recall	F-Measure	ROC
Spam	-	-	0.667	0.055	0.4	0.667	0.5	0.879
Non spam	-	-	0.945	0.333	0.981	0.945	0.963	0.879
All	93.1034%	6.8966%	-	-	-	-	-	-

6.3 Evaluating Arabic cloaking Web spam detection system

We used the same test dataset which consists of 5,000 Arabic Web pages to evaluate the Arabic cloaking Web spam Detection System; the results yields 94.1606% accuracy in detecting cloaking Web spam. Figure 6.5 shows the evaluation process of our Arabic cloaking Web spam detection system.

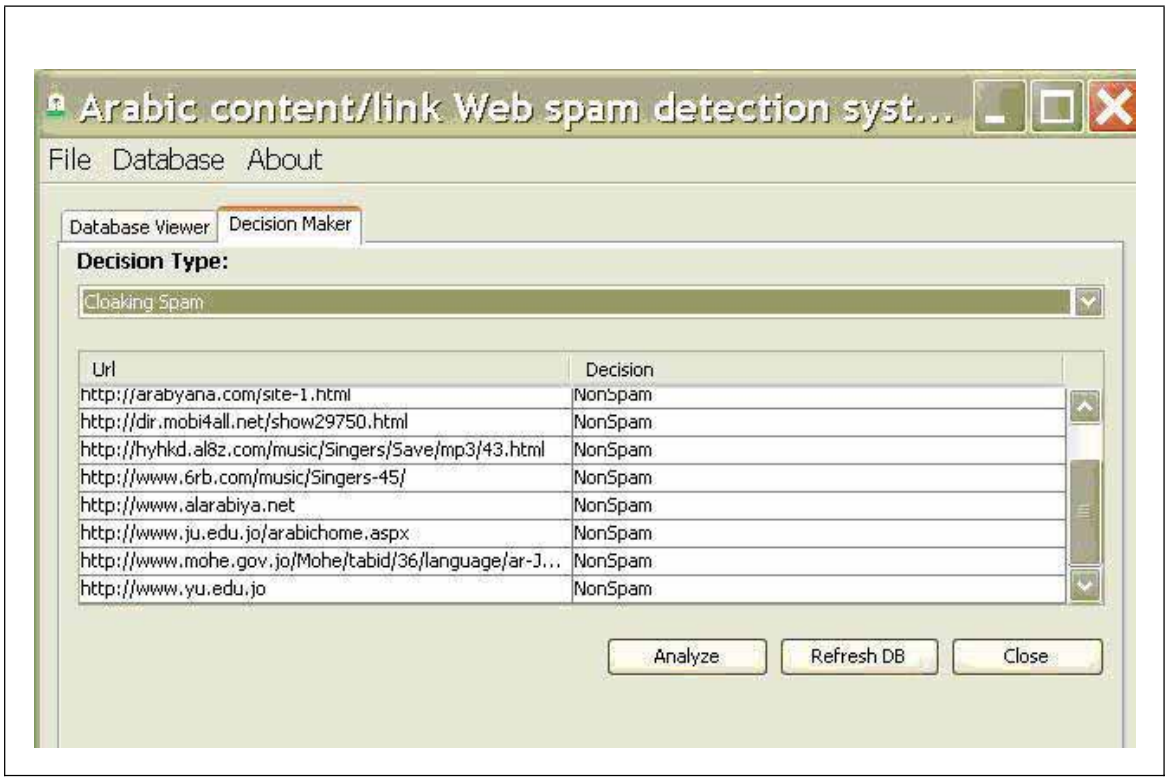


Figure 6.5: Evaluation process of Arabic cloaking Web spam detection system

The detailed evaluation results of Arabic Cloaking Web spam are shown in Table 6.3.

Table 6.3. Evaluation results of Arabic cloaking Web spam detectin system

Test Dataset	Accuracy	Error	True Positive Rate	False Positive Rate	Precision	Recall	F-Measure	ROC
Spam	-	-	0.929	0.045	0.956	0.929	0.942	0.995
Non spam	-	-	0.995	0.071	0.928	0.955	0.941	0.995
All	94.1606%	5.8394%	-	-	-	-	-	-

6.4 Evaluating Arabic content/link Web spam detection system

Finally we used the same test dataset which consists of 5,000 Arabic Web pages to evaluate our Arabic content/link Web spam Detection System. Where the system yields 89.011% accuracy in detecting content/link Web spam. The full results are shown in Table 6.4.

Table 6.4. Evaluation Arabic content/link Web spam results

Test Dataset	Accuracy	Error	True Positive Rate	False Positive Rate	Precision	Recall	F-Measure	ROC
Spam	-	-	0.966	0.55	0.911	0.966	0.938	0.9
Non spam	-	-	0.45	0.034	0.692	0.45	0.544	0.9
All	89.011 %	10.989%	-	-	-	-	-	-

6.5 Comparision between all types of Arabic content/link Web spam detection system

From the above subsections, we found that the Arabic cloaking detection system yields more accurate results than the other systems. Because the Arabic cloaking detection system monitors all cloaking features in the Web pages. Table 6.5 shows the comparisons of the Accuracy values between all types of spam in our Arabic content/link Web spam Detection System.

Table 6.5. Comparison between the accuracy values for all spam types

Test Dataset	True Positive Rate	False Positive Rate	Precision	Recall	F-measure	ROC
Content-based	0.901	0.452	0.893	0.901	0.891	0.882
Link-based	0.931	0.139	0.951	0.931	0.939	0.879
Cloaking	0.942	0.058	0.942	0.942	0.942	0.955
Content/link	0.89	0.47	0.87	0.89	0.88	0.9

Table 6.5 shows the superiority of the Arabic cloaking Web spam detection system relative to others. Followed by link-based, then content-based, and the content/link Arabic Web spam detection system respectively.

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

1. Kappa statistic (KS): is the statistical measure that explains the reduction in errors compared to the errors of a completely classification random (Witten & Frank, 2005).
2. Mean Absolute Error (MAE): the mean absolute error measures how the predictions are close to the actual outcomes (Witten & Frank, 2005).
3. Root Mean Squared Error (RMSE): is a measure of the differences between estimated values and actual values. It is related to the error variance or standard deviation. If RMSE is closer to zero, the prediction is considered good (Al-Kabi, *et al*, 2012).
4. Root Absolute Error (RAE): is the error prediction which presents a percentage error of a simple prediction model (Al-Kabi, *et al*, 2012).
5. Root Relative Squared Error (RRSE): is relative to what it would have been if a simple predictor had been used. It is obtained by taking the square root of the Relative squared error (Al-Kabi, *et al*, 2012).

Table 6.6 presents the comparisons of the performance measurements for all Arabic Web spam types.

Table 6.6. Performance measurements for all Arabic Web spam types

Types of Arabic Web spam	KS	MAE	RMSE	RAE	RRSE
Content-based	0.5319	0.1763	0.2856	62.8516%	80.1697%
Link-based	0.4654	0.0946	0.2586	72.6001%	115.2463%
Cloaking	0.8832	0.0527	0.2188	10.5366%	43.7482%
Content/link	0.4861	0.1639	0.2785	58.4385%	78.1673%

Table 6.6 presents clearly the effectiveness of our Arabic content/link Web spam Detection System in detecting all Arabic Web spam types, cloaking type in particular with the highest values for all Performance measurements used.

CHAPTER SEVEN

CONCLUSIONS AND FUTURE WORK

This chapter is divided into two sections. The first represents the conclusions of this thesis, while the second section represents the future work.

7.1 Conclusions

The continuous expansion of the Internet, lead to increase in the number of challenges to Web search companies to offer relevant and high quality Arabic information to its Arab users. So to accomplish their goals the Web sites owners attempt to adopt legal and illegal ways to lets their Web pages rank higher than they deserve in the SERP, to gain more users, and more revenues.

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

Three main Web spam types were studied in this thesis: Arabic content-based, Arabic link-based, and Arabic cloaking Web spam. The main goal of this study is to solve the Arabic Web spam detection problem. We discussed the relation between the Arabic Web spam types. In this thesis a larger Arabic content/link based spam dataset relative to those used in our previous studies was built and used. This large dataset

contains 28,000 Arabic spam Web pages which were collected through an enhanced embedded Web crawler. The spam dataset is divided into two parts; training dataset which consists of 23,000 Web pages used to build Arabic content/link Web spam detection system, and test dataset which consists of 5,000 Web pages used to evaluate Arabic content/link Web spam detection system. We built Arabic Web spam analyzer which extracts a larger number of features for content-based, link-based, and cloaking features.

The extracted features used by three classification algorithms to identify the best algorithm to detect the Arabic content/link Web spam. The rules of Decision Tree were extracted with 2% percentage group spam dataset, which is considered as the best algorithm to detect the Arabic content/link Web spam. Then we evaluated the Arabic content/link Web spam detection system, using test dataset that contains around 5,000 Web pages, and we gained good results for all Arabic Web spam types.

This thesis presented novel Arabic content/link Web spam detection system, which capable to detect three main types of Arabic Web spam. The system was implemented and tested on the Arabic spam dataset, and yields good results that would save the time of Arabic users, efforts, and help to retrieve the relevant results that satisfy their needs.

The proposed Arabic Web spam detection system is characterize by its flexibility, since we can increase the number of Web pages in the spam dataset, using the same methodology steps, and extract the new rules of the best Arabic Web spam detection algorithm. Then used the test dataset to evaluate the efficiency of the new rules.

7.2 Future work

We plan to extend this work in the future to include the following three areas:

1. Enhancing this work by including all the challenges of spam factors that influence the reputable Web pages and link popularities.

2. Detecting the spam techniques in the social networks, such as: FaceBook, YouTube, Google + and Twitter. As they are attracting more and more internet users, and they are targets for spammers.

3. Studying and investigating the detection of the malicious links in Arabic spammed Web pages. Malicious links usually combines between Web spam techniques and Web security issues particularly malware types (Worms and Viruses).

REFERENCES

Abernethy, J., Chapelle, O. and Castillo, C. 2008. *Web spam Identification Through Content and Hyperlinks*. In Proceedings of the Fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '08), Beijing, China. Pp 41-44.

Al-Kabi, M., Wahsheh, H., AlEroud, A. and Alsmadi, I. 2011. *Combating Arabic Web spam Using Content Analysis*. In Proceedings of the 2011 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT2011), Amman Jordan. Pp 401-404.

Al-Kabi, M., Wahsheh, H., Alsmadi, I., Al-Shawakfa, E., Wahbeh, A. and Al-Hmoud, A. 2012. Content Based Analysis to Detect Arabic Web spam. *Journal of Information Science. 38* (3): 284-296.

Almeida, R., Mozafar, B. and Ch, J. 2007. *On the Evolution of Wikipedia*. In Proceedings of the International Conference on Weblogs and Social Media. Boulder, Colorado, USA. Pp 1-8.

Alrawabdeh, W. 2009. Internet and the Arab World: Understanding the Key Issues and Overcoming the Barriers. *The International Arab Journal of Information Technology. 6* (1): 27-33.

Araujo, L. and Martinez-Romo, J. 2010. Web spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models. *IEEE Transactions on Information Forensics and Security. 5* (3): 581-590.

Attia, M. 2011. *Arabic Language Research and Translation*. Retrieved March, 23, 2012 from the World Wide Web: http://attiaspace.com

Baeza-Yates, R. and Ribeiro-Neto, B. 2010. *Modern Information Retrieval: The Concepts and Technology behind Search*, Addison-Wesley Professional, Indianapolis, Indiana, USA.

Batzios, A., Dimou, C., Symeonidis, A. and Mitkas, P. 2008. BioCrawler: An intelligent crawler for the semantic Web. *Expert Systems with Applications. 35*, 524–530.

Becchetti, L., Castillo, C., Donato, D., Leonardi, S. and Baeza-Yates, R. 2008. *Web spam Detection: Link-based and Content-based Techniques*. In The European Integrated Project Dynamically Evolving, Large Scale Information Systems (DELIS): proceedings of the final workshop, Barcelona, Spain. *222*, Pp 99-113.

Becchetti, L., Castillo, C., Donato, D., Leonardi, S. and Baeza-Yates, R. 2006a. *Link-Based Characterization and Detection of Web spam*. In Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web (AIRWEB'06), Seattle, Washington, USA. Pp 1-8.

Becchetti, L., Castillo, C., Donato, D., Leonardi, S. and Baeza-Yates, R. 2006b. *Using rank propagation and probabilistic counting for link-based spam detection*. In Proceedings of the Workshop on Web Mining and Web Usage Analysis (WebKDD), ACM, Philadelphia, Pennsylvania, USA. Pp 1-10.

Benczur, A., Biro, I., Csalogany, K. and Sarlos, T. 2007. *Web spam Detection via Commercial Intent Analysis*. In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web Pages (AIRWeb'07), ACM, Banff, Alberta, Canada. Pp 89-92.

Benczur, A. A., Siklosi, D., Szabo, J., Biro, I., Fekete, Z., Kurucz, M., Pereszlenyi, A., Racz, S. and Szabo, A. 2008. *Web spam: a Survey with Vision for the Archivist*. International Web Archiving Workshop (IWAW'08), Aaarhus, Denmark. Pp. 1-9.

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

Berlt, K., Moura, E. S., Carvalho, A., Cristo, M., Ziviani, N. and Couto, T. 2010. Modeling the web as a hypergraph to compute page reputation. *Information Systems. 35* (5): 530-543.

Beseiso, M., Ahmad, A. and Ismail, R. 2010. A Survey of Arabic Language Support in Semantic Web. *International Journal of Computer Applications. 9*(1): 35–40.

Biggio, B., Fumera, G., Pillai, I. and Roli, F. 2011. A survey and experimental evaluation of image spam filtering techniques. *Pattern Recognition Letters. 32*, 1436–1446.

Boone, G., Secci, J. and Gallant, L. 2009. Emerging Trends in Online Advertising. *doxa comunicacion. 5* (5): 241-253.

Castillo, C., Corsi, C. and Donato, D. 2008. *Query-log mining for detecting spam*. In Proceedings of the 4th international workshop on Adversarial information retrieval on the web Pages (AIRWeb '08), ACM, Beijing, China. Pp 17-20.

Castillo, C., Donato, D., Gionis, A., Murdock, V. and Silvestri, F. 2007. *Know your neighbors: Web spam detection using the Web topology*. In Proceedings of the

30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, Netherlands. Pp 423-430.

Caverlee, J. and Liu, L. 2007. *Countering Web spam with Credibility-Based Link Analysis*. In Proceedings of the annual ACM Symposium on principles of Distributed Computing, Portland, Oregon, USA. 26, Pp. 157-166.

Chau, M. and Chen, H. 2008. A machine learning approach to web page filtering using content and structure analysis. *Decision Support Systems. 44* (2): 482-494.

Chellapilla, K. and Chickering, D. M. 2006. *Improving Cloaking Detection Using Search Query Popularity and Monetizability*. In Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 06), Seattle, Washington, USA. Pp 17-24.

Chellapilla, K. and Maykov, A. 2007. *A taxonomy of JavaScript redirection spam*. In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web (AIRWeb '07), ACM, Banff, Alberta, Canada. Pp 81-88.

Chung, Y., Toyoda, M. and Kitsuregawa, M. 2010. *Identifying spam link generators for monitoring emerging web spam*. In Proceedings of the 4th workshop on Information credibility (WICOW '10), Raleigh, North Carolina, USA. Pp 51-58.

Chung, Y., Toyoda, M. and Kitsuregawa, M. 2009. *Detecting Link Hijacking by Web spammers*. In Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD '09), ACM, Berlin, Heidelberg. Pp 339-350.

Dai, N., Davison, B. D. and Qi, X. 2009. *Looking into the Past to Better Classify Web spam*. Fifth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '09), Madrid, Spain. Pp 1-8.

Dimou, C., Batzios, A., Symeonidis, A. and Mitkas, P. 2006. *A Multi-Agent Simulation Framework for Spiders Traversing the Semantic Web*. In Proceedings of International Conference on Web Intelligence, Hong Kong, China. Pp 736-739.

Dou, W., Lim, K. H., Su, C., Zhou, N. and Cui, N. 2010. Brand Positioning Strategy Using Search Engine Marketing. *MIS Quarterly. 34* (2): 261-279.

Du, Y., Shi, Y. and Zhao, X. 2007. *Using spam farm to boost PageRank*. In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web (AIRWeb '07), ACM, Banff, Alberta, Canada. Pp 29-36.

Erdelyi, M., Benczur, A. A. 2011b. *Temporal Analysis for Web spam Detection: An Overview*. In proceedings of the 1st International Temporal Web Analytics Workshop (TWAW 2011), Hyderabad, India. Pp 17-24.

Erdelyi, M., Garzo, A. and Benczur A. A. 2011a. *Web spam Classification: a Few Features Worth More*. In Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality WebQuality '11, ACM, Hyderabad, India. Pp 27-34.

Ermakova, L. 2011. *Transforming Message Detection*. Young Scientists Conference in Information Retrieval, Voronezh, Russian. Pp 15-29.

Egele, M., Kolbitsch, C. and Platzer, C. 2011. Removing web spam links from search engine results. *Journal in Computer Virology. 7*(1): 51-62.

Fetterly, D., Manasse, M. and Najork, M. 2004. *spam, damn spam, and statistics: using statistical analysis to locate spam Web pages*. In Proceedings of International Workshop on the Web and Databases (WebDB '04), Paris, France. Pp 1-6.

Fetterly, D., Manasse, M. and Najork, M. 2005. *Detecting phrase-level duplication on the World Wide Web*. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil. Pp 170–177.

Gadge, J., Sane, S. S. and Kekre, H. B. 2011. *Layered Approach to Improve Web Information Retrieval*. In Proceedings of 2nd National Conference on Information and Communication Technology (NCICT), Nagpur, India. 7, Pp 28-32.

Geng, G., Li, Q. and Zhang, X. 2009. *Link based small sample learning for web spam detection*. In Proceedings of the 18th international conference on World wide web WWW '09, ACM, Madrid, Spain. Pp 1185-1186.

Geng, G., Wang, C., Li, Q., Xu, L. and Jin, X. 2007. *Boosting the Performance of Web spam Detection with Ensemble Under-Sampling Classification*. Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), Haikou, Hainan, China. Pp 583-587.

Goodstein, M. and Vassilevska, V. 2007. *A Two Player Game To Combat Web spam*. School of Computer Science, Carnegie Mellon University, Pittsburgh, USA.

Gyongyi, Z. and Garcia-Molina, H. 2005. *Web spam taxonomy*. In Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web, Chiba, Japan, Pp. 1-9.

Gyongyi, Z., Garcia-Molina, H. and Pedersen, J. 2004. *Combating Web spam with TrustRank*. In Proceedings of the 30th International Conference on Very Large Databases (VLDB), Toronto, Canada. 30, Pp 576-587.

Hayati, P., Chai, K., Potdar, V. and Talevski, A. 2010a. *Behavior-Based Web spambot Detection by Utilising Action Time and Action Frequency*. In Proceedings of International Conference for Computational Science and its Applications (ICCSA), Fukuoka, Japan. 2, Pp 351-360.

Hayati, P., Potdar, V., Chai, K. and Talevski, A. 2010b. *Web spambot Detection Based on Web Navigation Behavior*. In Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications, Perth, Australia. Pp 797-803.

Hayati, P. and Potdar, V. 2009. *Toward spam 2.0: An Evaluation of Web 2.0 Anti-spam Methods*. 7th IEEE International Conference on Industrial Informatics (INDIN 2009), Cardiff, Wales. Pp 875-880.

Hayati, P., Chai, K., Potdar, V. and Talevski, A. 2009. *Honeyspam 2.0: Profiling Web spambot Behavior*. In Proceedings of the 12th International Conference on Principles of Practice in Multi-Agent Systems (PRIMA '09), Nagoya, Japan. Pp 335-344.

Hochman, J. 2012. *How to Get More Pages into Google's Index*. Retrieved April, 25, 2012 from the World Wide Web:

http://searchenginewatch.com/article/2048484/How-to-Get-More-Pages-into-Googles-Index

Jaramh, R., Saleh, T., Khattab,S. and Farag, I. 2011. Detecting Arabic spam Web pages using Content Analysis. *International Journal of Reviews in Computing*. 6, 1-8.

Jayanthi, S. K. and Sasikala, S. 2011. DBLC_SPAMCLUST: spamdexing Detection By Clustering Clique-Attacks In Web Search Engines. *International Journal of Engineering Science and Technology (IJEST)*. 3 (6): 4572- 4580.

Jones, T., Hawking, D. and Sankaranarayana, R. 2007. *A Framework for Measuring the Impact of Web spam*. In Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia. Pp 108-111.

Kang, F., Liu, X., and Liu, W. 2011. *APersonalized Ranking Approach via Incorporating Users' Click Link Information into PageRank Algoritm*. In Proceedings of the 2011 International Conference on Energy Systems and Electrical Power (ESEP 2011), Singapore. 13, 275-284.

Kerchove, C., Ninove, L. and Dooren, P. 2008. Maximizing PageRank via external links. *Linear Algebra and its Applications. 429*, 1254–1276.

Kolari, P., Java, A., Finin, T., Oates, T. and Joshi, A. 2006. *Detecting spam blogs: a machine learning approach*. In Proceedings of the 21st national conference on Artificial intelligence (AAAI'06), Boston, Massachusetts. 2, Pp1351-1356.

Kumar, E. and Kohli, S. 2011. *Improving Link spam Detection using spamizer*, In Proceedings of the World Congress on Engineering and Computer Science 2011 (WCECS 2011), San Francisco, USA. 1, Pp 19-21.

Piskorski, J., Sydow, M. and Weiss, D. 2008. Exploring *Linguistic Features for Web spam Detection*. Fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 08), Beijing, China. Pp 1-4.

Largillier, T. and Peyronnet, S. 2011. Detecting Webspam Beneficiaries Using Information Collected by the Random Surfer. *International Journal of Organizational and Collective Intelligence IJOCI. 2* (2): 1-17.

Liang, C., Ru, L. and Zhu, X. 2007. R-spamRank: A spam detection algorithm based on link analysis. *Journal of Computational Information Systems. 3*, 1705-1712.

Lin, J. 2009. Detection of cloaked Web spam by using tag-based methods. *Expert Systems with Applications. 36*, 7493-7499.

Liu, T., Xu, J., Qin, T., Xu, J., Xiong, W. and Li, H. 2007. *LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval*. SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007), Amsterdam, Netherlands. Pp 1-10.

Martinez-Romo, J. and Araujo, L 2012. Updating broken web links: An automatic recommendation system. *Information Processing and Management. 48*, 183–203.

Martinez-Romo, J. and Araujo, L. 2009. W*eb spam Identification Through Language Model Analysis*. Fifth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '09), Madrid, Spain. Pp 21-28.

Mohammad, A. H. and Zitar, R. A. 2011. Application of genetic optimized artificial immune system and neural networks in spam detection. *Applied Soft Computing. 11*, 3827–3845.

Najadat, H. and Hmeidi, I. Web spam Detection Using Machine Learning in Specific Domain Features. *Journal of Information Assurance and Security. 3*, 220-229.

Narisawa, K., Bannai, H., Hatano, K. and Takeda, M. 2007. *Unsupervised spam Detection based on String Alienness Measures*. In Proceedings of the 10th international conference on Discovery science Pages (DS'07), ACM, Banff, Canada. Pp 161-172.

Ntoulas, A., Najork, M., Manasse, M. and Fetterly, D. 2006. *Detecting spam Web Pages through Content Analysis*. In Proceedings of the World Wide Web Conference, Edinburgh, Scotland. Pp 83–92.

Niu, X., Ma, J., He, Q., Wang, S. and Zhang, D. 2010. *Learning to Detect Web spam by Genetic Programming*. In Proceedings of the 11th international conference on Web-age information management (WAIM'10), Jiuzhaigou, China. Pp 18–27.

Niu, Y., Wang, Y., Chen, H., Ma, M. and Hsu, F. 2006. *A Quantitative Study of Forum spamming Using Context-based Analysis*. In Proceedings of the Network & Distributed System Security (NDSS) Symposium, San Diego, California, USA.Pp. 1-14.

Turdakov, D. and Simanovsky, A. 2011. *Detecting Content spam on the Web through Text Diversity Analysis*. In Proceedings of the 7th Spring Researchers Colloquium on Databases and Information Systems (SYRCoDIS), Moscow, Russia. Pp 277-296.

Ryding, K 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press, New York, USA.

Saraswathi, D., Vijaya Kathiravan, A. and Anita, S. 2011. A Novel Approach for Combating spamdexing in Web using UCINET and SVM Light Tool.

International Journal of Innovative Technology and Creative Engineering. 1 (3): 47- 52.

Selvan, M. P., Sekar, A. C. and Dharshini, A. P. 2012. Survey on Web Page Ranking Algorithms. *International Journal of Computer Applications. 41* (19):1-7.

Shen, G., Gao, B., Liu, T., Feng, G., Song, S. and Li, H. 2006. *Detecting Link spam using Temporal Information*. In Proceedings of the Sixth International Conference on Data Mining Pages (ICDM '06), IEEE, Hong Kong, China. Pp 1049-1053.

Spirin, N. and Han, J. Survey on Web spam Detection: Principles and Algorithms. *SIGKDD Exploration. 13* (2): 50-64.

Svore, K.M., Wu, Q., Burges, C. J. C. and Raman, A. 2007. *Improving Web spam Classification using Rank-time Features*. In Proceedings of Adversarial Information Retrieval on the Web (AIRWeb'07), Banff, Alberta, Canada. Pp 1-8.

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

Tian, Y., Weiss, G. M. and Ma, Q. 2007. *A Semi-Supervised Approach for Web spam Detection using Combinatorial Feature-Fusion*. In Proceedings of the Graph Labeling Workshop and Web spam Challenge (GRAPHLAB 2007), Warsaw, Poland. Pp 16-23.

Vangapandu, K., Brewer, D. and Li, K. 2009. *A Study of URL Redirection Indicating spam*. In Proceedings of the Fourth Conference on Email and Anti-spam (CEAS 2009), California USA. Pp 1-9.

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

Wahsheh, H. A., Abu Dosh, I., Al-Kabi, M., Alsmadi, I. and Al-Shawakfa, E. 2012a. Machine Learning Algorithms to Detect Content-based Arabic Web spam. *Journal of Information Assurance and Security. 7* (1): 14-24.

Wahsheh, H. A., Al-Kabi, M. N. and Alsmadi, I. M. 2012b. *spam Detection Methods for Arabic Web Pages*. First Taibah University International Conference on Computing and Information Technology (ICCIT 2012), Al-Madinah Al-Munawwarah, Saudi Arabia. *2*, Pp 486-490.

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

Wahsheh, H., Alsmadi, I. and Al-Kabi, M. 2012d. Analyzing the Popular Words to Evaluate spam in Arabic Web Pages. *IJJ: The Research Bulletin of JORDAN ACM – ISWSA. 2* (2): 22-26.

Wang, Y. 2005. A multinomial logistic regression modeling approach for anomaly intrusion detection. *Computers & Security. 24*, 662-674.

Wang, Y., Ma, M., Niu, Y. and Chen, H. 2007a. *spam Double-Funnel: Connecting Web spammers with Advertisers*. International World Wide Web Conference Committee (IW3C2), Banff, Alberta, Canada. Pp 8–12.

Wang, W., Zeng, G., Sun, M., Gu, H. and Zhang, Q. 2007b. EviRank: An Evidence Based Content Trust Model for Web spam Detection. *In: Advances in Web and Network Technologies, and Information Management,* (Chang, K.C. Ed.) (APWeb/WAIM 2007 Ws), pp. 299–307.

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

Wang, W., Zeng, G. and Tang D. 2010. Using evidence based content trust model for spam detection. *Expert Systems with Applications*. 37(8): 5599-5606.

Webb, S., Caverlee, J. and Pu, C. 2007. *Characterizing Web spam Using Content and HTTP Session Analysis*. In Proceedings of the Fourth Conference on Email and Anti-spam (CEAS 2007), Mountain View, California, USA. Pp 1-9.

West, A., Agrawal, A., Baker, P., Exline, B. and Lee, I. 2011. *Autonomous link spam detection in purely collaborative environments*. In Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11), ACM, Mountain View, California, USA. Pp 91-100.

Wu, B. and Davison, B. D. 2005. *Cloaking and Redirection: A Preliminary Study*. In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'05), Chiba, Japan. Pp 1-10.

Witten I. H. and Frank, E. 2005. Data Mining: Practica Machine Learning Tools and Techniques, *Morgan Kaufmann Series in Data Management Systems*, second edition, Morgan Kaufmann (MK).

Xhemali, D., Hinde, C. J. and Stone, R. G. 2009. Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *International Journal of Computer Science. 4*, 6-23.

Yang, L. 2006. *Distance Metric Learning: A Comprehensive Survey*. Department of Computer Science and Engineering Michigan State University.

Yoon, J. W., Kim, H. and Huh, J. H. 2010. Hybrid spam filtering for mobile communication. *Computers & security.29*, 446–459.

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

Zhang, W., Zhu, D., Zhang, Y., Zhou, G. and Xu, B. 2011. *Harmonic functions based semi-supervised learning for Web spam detection*. In Proceedings of ACM Symposium on Applied Computing, Taichung, Taiwan. Pp 74-75.

Zhou, D., Burges, C. and Tao, T. 2007. *Transductive Link spam Detection*. In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web Pages (AIRWeb'07), ACM, Banff, Alberta, Canada. Pp 1-8.

Zhou, B. and Pei, J. 2009. Link spam target detection using page farms. *ACM Transactions on Knowledge Discovery from Data (TKDD). 3* (3): 1-38.

Zhu, L., Sun, A. and Choi, B. 2011. Detecting spam blogs from blog search results. *Information Processing and Management: an International Journal. 47* (2): 246-262.

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

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

Spamlaws, 2012, *spam Statistics and Facts*. Retreived February, 24, 2012 from the World Wide Web: http://www.spamlaws.com/spam-stats.html

Wikipedia, 2012. *List of countries where Arabic is an official language*. Retreived February, 24, 2012 from the World Wide Web: http://en.wikipedia.org/wiki/List_of_countries_where_Arabic_is_an_official_language

الخلاصة

نحو بناء نظام لاكتشاف الصفحات العربية غير المرغوب بها

إعداد
حيدر أحمد وحشة

إشراف
د. محمد ناجي الكعبي

يتميز الإنترنت بديناميكيته، إذ يضاف له عدد متزايد من صفحات الإنترنت، وهذا يؤثر نسبيا على إجمالي حجم محتويات الإنترنت. تعتبر محركات البحث (Search Engines) اليوم واحدة من البوابات الأساسية المستخدمة لولوج عالم الإنترنت، وللحصول على الأنواع المختلفة من المعلومات والوثائق والملفات. ويتلخص الهدف الأساسي لمحركات البحث الجيدة بعرض أكبر نسبة من عناوين صفحات الإنترنت بناء على الكلمات أو الاستعلامات من المستخدمين. تتصف محركات البحث الجيدة بميزتين أساسيتين هما السرعة وملائمة النتائج للكلمات أو العبارات التي بحث من خلالها المستخدم. إضافة إلى ضرورة أن يتصف محرك البحث بالشمولية التي تعني زيادة وإضافة كل الملفات الجديدة إلى فهرس محرك البحث، ليتم عرض هذه النتائج عند الضرورة.

هناك اليوم العديد من مالكي المواقع الذين يسعون إلى تضليل محركات البحث مستفيدين من الاستخدام غير الشرعي للتقنيات الواردة في ما يعرف بتحسين محركات البحث (Search Engine Optimization) والمعروف اختصارا بـ (SEO). ويهدف هؤلاء المالكون من وراء هذا الاستخدام غير الشرعي إلى جعل عناوين صفحات الإنترنت الخاصة بهم تظهر في أعلى قائمة النتائج (العناوين) وضمن النتائج العشرة الأوائل لمحركات البحث، لتحتل ضمن ترتيب النتائج موقعا متقدما لا تستحقه أصلا. ويُعزى سبب هذا التصرف إلى محاولة هؤلاء زيادة عدد الزائرين لصفحات الإنترنت الخاصة بهم التي عادة ما تحوي معلومات تتعلق بالتجارة والبيع. تمثل هذه الدراسة بناء أول نظام لاكتشاف صفحات الإنترنت العربية غير المرغوب بها (Arabic Web spam) والمعتمد في عمله على محتوى هذه الصفحات (Content) وارتباطاتها التشعبية (Links). يتميز النظام المقترح والجديد بفكرته وقُدراته على استخلاص مجموعة من خصائص محتويات صفحات الإنترنت وارتباطاتها التشعبية. و هَدُف هذا البحث بناء مجموعة بيانات (Dataset) كبيرة لصفحات الإنترنت العربية غير المرغوب بها، إذ تضم هذه المجموعة من البيانات ثلاثة مجموعات فرعية من بيانات صفحات الإنترنت غير المرغوب بها، و تبلغ النسب المئوية لهذه المجموعات الثلاثة الفرعية 2%، 30%، و 40% على التوالي. تم جَمع مجموعة البيانات لصفحات الإنترنت العربية غير المرغوبة، باستخدام زاحف الشبكة (Web Crawler)، والذي يمثل أحد المكونات الأساسية للنظام المقترح والذي تقوم عليه هذه الدراسة. ويقوم عمل النظام المقترح والذي تم بناؤه على قواعد شجرة القرار(Decision Tree) ، والتي تُعتبَر أفضل مُصنف لاكتشاف الصفحات غير المرغوب بها بالاعتماد

على محتويات صفحة الإنترنت وارتباطاتها التشعبية. يُوفر النظام المقترح حلولا لمشاكل صفحات الإنترنت العربية غير المرغوب بها، إذ يوفر الجهد والوقت لمستخدمي الإنترنت، ويُساعد هذا النظام على تنقية نتائج محركات البحث من جميع عناوين صفحات الإنترنت العربية غير المرغوب بها. أثبتت التجارب التي أجريت على النظام المقترح والمنشأ بأن نسبة دقته لاكتشاف صفحات الإنترنت العربية غير المرغوب بها اعتمادا على محتوى الصفحات قد بلغت 90.1099%، أما نسبة دقة النظام المذكور لاكتشاف صفحات الإنترنت العربية غير المرغوب بها اعتمادا على الارتباطات التشعبية قد بلغت 93.1034%، أما نسبة دقة النظام لاكتشاف صفحات الإنترنت العربية غير المرغوب بها اعتمادا على أسلوب التخفي (cloaking) قد بلغت 94.1606%، وأخيرا فان نسبة دقة النظام لاكتشاف صفحات الانترنت العربية غير المرغوب بها اعتمادا على محتوى الصفحات وارتباطاتها التشعبية قد بلغت 89.011%.

الكلمات المفتاحية: صفحات الإنترنت العربية غير المرغوب بها، محتوى صفحات الإنترنت، الارتباطات التشعبية، أسلوب التخفي.

إلى والديّ . . بكل ما يستطيع القلب من حب

إلى أستاذي الفاضل الدكتور محمد الكعبي

اعترافا بفضله وتقديرا لعلمه

وإشادة بجهوده الموفقة على الدوام

و إلى كل من سار معي في رحلتي . . إخوتي وأصدقائي

كل الشكر والتقدير

يرفع الله الذين آمنوا منكم والذين أوتوا العلم درجات والله بما تعملون خبير

المجادلة: ١١

نحو بناء نظام لاكتشاف الصفحات العربية غير المرغوب بها

إعداد

حيدر احمد وحشه

بكالوريوس نظم معلومات حاسوبية، جامعة اليرموك 2009

قدمت هذه الرسالة استكمالا لمتطلبات الحصول على درجة الماجستير في نظم المعلومات الحاسوبية من قسم نظم المعلومات الحاسوبية في كلية تكنولوجيا المعلومات وعلوم الحاسوب بجامعة اليرموك

موافق عليها من قبل:

د. محمد ناجي الكعبي.....(مشرفا رئيسا)
استاذ مساعد / قسم نظم المعلومات الحاسوبية / جامعة اليرموك.

د. بلال مصطفى ابوعطا .....(عضوا)
استاذ مساعد / قسم نظم المعلومات الحاسوبية / جامعة اليرموك.

د. عامر فضيل البدارنة .....(عضوا)
استاذ مشارك / قسم نظم المعلومات الحاسوبية / جامعة العلوم والتكنولوجيا الأردنية.

17/7/2012

نحو بناء نظام لاكتشاف الصفحات العربية غير المرغوب بها

رسالة ماجستير
في
نظم المعلومات الحاسوبية

إعداد
حيدر احمد وحشه

إشراف
د. محمد الكعبي

قسم نظم المعلومات الحاسوبية

جامعة اليرموك

17/7/2012