International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.3,No.4, 2013
Published online: April 25, 2013
DOI: 10.7321/jscse.v3.n4.1

Designing the Matrix of Analysis to Solve the Problems Regarding the Web Site Searching

MohammadReza DehghaniMahmoudAbadi *[1], Hassan Samavarchi[2]

[1] Islamic Azad University, Ashkzar, Yazd, Iran

[2]Yazd University, Iran

Email: [1]phdmrdma@gmail.com, [2]hsamavarchi@yazduni.ac.ir

Abstract: **The discovery of the useful patterns from Word Wide Web (WWW) known as a web searching is one of the basic uses of the data searching. Today's searching engine motors are no longer able to meet the needs of the users in the enormous libraries whether they are the web searchers or the gate ways of the web. This study focuses on better and quicker discovery of information from the World Wide Web. In this paper we offer a comprehensive structure of a data searching system, this study also investigates into the importance of the web searching and its problems facing the discovery of the useful information from the webs and their solutions in the forms of tables, named SPAM and T and A.**

Keywords: *data searching, web searching, searching motors, web gateway.*

* Corresponding Author:
MohammadReza DehghaniMahmoudAbadi,
Department of Computer Engineering, Islamic Azad University, Ashkzar,
Yazd, Iran.
Email: phdmrdma@gmail.com Tell: +983523625951 Fax: 983523625953

1. Introduction

We live in the age of information, the age in which humans produce more data and spread them more than ever. In fact, the information we receive is more than what we are able to analyses. So the user finds it increasingly difficult to get what he/she needs. The following are the reasons why there exists such a problem[1].

1) The web is so great and varied.
2) The data are increasingly changing
3) The web is badly organized.

Each user may be interested in one small part of the web. Therefore, the users are faced with many problems finding the desired information. The searching motors help the users obtain the related sources of information; however, they are faced with many problems. One title may involve hundreds or even thousands documents that is why the searching motor may return a lot of documents including the unrelated ones. A lot of seeming related documents may not contain the key words describing and defining that particular subject[2].

The collection of information available in the web can be regarded as a rich resource for the web searching, assisting us in finding the sources of information by improving the function of searching motors. This article is intended to briefly study the data searching and its procedure, offering a comprehensive structure based on them. This study also focuses on the position of web searching in data searching and studies the problems involved in the efficient discovery of information from the web by using SPAM and T and A.

SPAM is a tool which provides us with a framework needed to choose an appropriate method for creating a web searching system. T and A is tool used for selecting the tools and appropriate ways needed for constructing a web searching system. These tools have been tested and evaluated in a system called Abut Universities Portal a gateway which makes access to the classified collections of different university sites across the globe[3].

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.3,No.4, 2013
Published online: April 25, 2013
DOI: 10.7321/jscse.v3.n4.1

Finally, the conclusions of the studies aiming to utilize the data searching techniques, especially the web searching methods in the searching motors are offered. The Data searching You are so deemed to get off on the wrong foot. Data searching, defined as the discovering of science from the data bases, is a procedure for mining the efficient patterns from the data bases.

2. Data Mining

The data mining can mine the desired patterns for users from a variety of data bases. Most researchers consider data mining synonymous with the discovery of science from the data bases[4].

The discovery of science is made through the following procedure one after another:

1) Purifying the data.

2) Making the data uniform.

3) Transforming the data.

4) Mining data: It is a necessary Procedure in which the intelligent methods of the suitable data patterns.

5) Evaluating the pattern: The patterns minds are evaluated.

6) Showing and revealing the science: The different techniques for showing. Sciences to show the users the science discovered.

2.1. The Structure of data mining system

Different structures for data mining systems are offered. Generally speaking, the structure of a data mining system consists of the following parameters[5]:

Data base, the analytical data base or another information store, like specific systems such as AGC4ISR[17] or in a cloud computing system that recognized by cloud template[15]. This parameter includes one or more databases, the analytical database or other information stores on which the techniques for deletion and uniformity are performed.

The servers of the database: They are responsible for providing the data based on the user's request.

The base of science: This is the store or the range of science which is used to assist in searching or evaluating the patterns Used.

The data mining motor includes some operational parameters for determining features, communal rules, and defining the structures of data through clustering and analytical evaluation.

The parameter for evaluating patterns: This parameter studies the relevancy of the patters and is connected with other data mining parameters so that searching for the desired patterns will fare well.

The user's graphic interface: this parameter makes it possible for the data bases to be reviewed, to search for the patterns and evaluate the patterns discovered.

3. Web mining

The web mining is one of the most important usages of the data mining in the World Wide Web to find and obtain the useful patterns. Not only is the web a mine of information, but it also contains a collection of links to have access to web pages and use the information which is a rich source of data mining.

3.1. A variety of operations done in the web mining.

Generally speaking, the desired operations in web mining include searching for the contents, the structures and the usages. It should be noted that these different branches are not quite disconnected from each other. These operations are employed to solve the problems caused when users search in the World Wide Web. To determine the suitable operation for solving the problems of the searching motors, Table 1 entitled as SPAM will be of value.

This table assists you with the problems encountered, the usages of the web mining, as well as the methods obtained from different articles. It is easy to get the desired operations to face the problems by using the table and its classification. The blank space in the table indicates that there is no method

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.3,No.4, 2013
Published online: April 25, 2013
DOI: 10.7321/jscse.v3.n4.1

related to that particular operation. The way of performance, and the results there of will be discussed here too[6].

	Searching the contents in the web	**Searching the structures in the web**	**Searching the usages**
Accessing the desired information through the searching motors in spite of the fact that the web is great and dynamic	Searching the contacts, dividing the documents in the web, relocating the information		
Enhancing the quality of internet services		The discovery of the credits for pages through the links of each page and hub	Searching the reports of access to the web servers and meta-information
The user can see part of the web which belongs to him/her	The classification of the documents in the web		Searching the reports of access to the web servers and meta-information
			Searching the reports of access to the web servers and meta-information
Giving web pages credits		The discovery of the credits for pages through the links of each page and hub	
The appropriate organization of the web	Classification of the web pages	Dividing the web through studying the structures of links among pages	Searching the reports of access to the web servers and meta-information

Table 1: SPAM choosing the desired web-mining operation to solve the problem caused in the path of discovering information from the World Wide Web.

3.1.1. The content-mining of the web

The content-mining of the web is defined as the discovery of the useful information obtained from documents and the contents of the web. What is the content of the web may include a variety of data. Internet used to offer a variety of services and resources/however, the web provides us with most of these data these days. Lately the government information has been developing extensively. The content of the web consist of unstructured data like free texts, semi-structured data like HTML, and finally a more structured data like the data in the table or the data bases producing HTML pages[7].

Most of the contents are unstructured. The investigation into the data-mining techniques in unstructured tax is called text mining. So the text-mining 8 can be considered as a type of content-mining of the text.

We can examine the studies conducted in the files of content-mining of the web from the two different points of view. the view of relocating information and that of data bases.

From the view of relocating the information, the content-mining is intended to improve the procedure of finding information or fulfilling the user's needs which is usually based on the profiles deducted or adapted for the users. On the other hand, from the view of data bases, the content-mining is basically intended to make a model out of data in the web and use them in a way that we can have access to base on the key words[8].

3.1.2. The structure- mining of the web

The structure- mining of the web is intended to discover a model underlying the links in the web. This model is based on the topology of the super links, it can be used for dividing the web pages, and generate information such as the similarity and connection among the web sites. The structure-mining can be used to discover the web page credits. Determining the pages credits is hidden in the links among the pages. Web is not a series of separate pages/its prominent feature is the links used to connect the pages. These links contain a lot of hidden information which help create the idea that credit evaluation of the pages is done automatically[9].

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.3,No.4, 2013
Published online: April 25, 2013
DOI: 10.7321/jscse.v3.n4.1

In fact, when the writer of a page puts a link with the next page. On his this link stands for a seal of approval the uses. The more links a page carries the more important the page will be and it will take more credits.

So the links provide comprehensive and complete information regarding the relevancy quality and the structure of the web it is also a rich source of web-mining. A lot of studies have been carried out concerning searching the reliable pages. In the 1970's, to evaluate the articles, the researchers introduced methods based on the ideas presented in the prestigious corresponding magazines.

In spite of the ideas presented in the journals, the web links have got special features whose use brings out problems for the credits of the pages. Each link doesn't indicate the approval we are looking for.

In fact some of the links are designed for different purposes. Second, one particular page may not have a link on the part of the rival writers, which reduces the credit of the pages due to the commercial or competitive reasons. For example, the Coca-Cola Company prefers not to approve Pepsi, its rival company by linking with its web pages. Third, the reliable pages don't describe themselves[10].

3.1.3. the usage-mining of the web

In the usage-mining of the web, the web's record reports are examined for discovering the user's access patterns. While content-mining and structure-mining of the web make use of the original and real data, the usage-mining of the web uses the secondary data derived from the user's interactions.

The practical information of the web includes the reports of the access to the web servers, the reports of the proxy servers, the reports of the researchers, The users' profiles, the users' interactions and meetings, users' questions, users' mailing list, the clicks of mouse's, or any kind of data regarded as the users' activities. The order observed in the records and its analysis can determine the prospective clients of e-commerce enhancing the quality of internet services and finally improving the efficiency of the web server[11].

4. Methods and wares used for the web-mining

To perform the various operations different methods and wares have been used. By using the SPAM table, the function of each web-mining operation and the solution to problems of obtaining information web. Methods and wares ideally suited for each activity and operation are provided in a table 2 entitled T and A. This table introduces the methods and tools needed for the web-mining.

The methods of web-mining		
Methods of and wares to the web-mining	**Techniques**	**Use**
The methods of text-mining	Episodic rules, processing method of natural language, the wares to the nervous system, clustering wares	Searching the unstructured text to improve the searching motor
The methods of relocating information	The hidden model of Marco, self-organizing maps, inductive methods	Searching the contents of the document in the web to improve the efficiency of searching motors, finding patterns in the contents
The methods of dividing the text	Rocchio, knearest neighbor, support vector machine, inductive methods, the excellent system, nervous system, statistical learning	Searching the contents of the document in the web to improve the efficiency of searching motor
Searching the communal rules	The communal rules	Identifying the schemes of web site to organize the web
The methods of classification	Decision-making tree-KNN- genetic algorithm, the neural network	The classification of the document in the web
The clustering method	The hierarchical methods, accumulative methods, the network methods	Clustering the documents, data mining the initial parts for division
The methods of statistical learning	Regression the little linear square	Dividing the documents in the web
The methods of statistical learning	Regression the little linear square	Dividing the documents in the web,

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.3,No.4, 2013
Published online: April 25, 2013
DOI: 10.7321/jscse.v3.n4.1

		changing the user into a model
Neural network	The backward publishing network	Dividing the documents in the web
Excellent network	Med index, construe	Dividing the documents in the web, changing the user into a model

Table 2: Table of T and A showing the wares and methods of web-mining

The techniques for each method have been offered. By using the table, a suitable method can be used based on the desired use and operational field. To employ the appropriate technique from the techniques introduced an evaluative bench-mark. The primary benchmarks for the evaluation will be discussed[12].

5. The primary benchmarks for comparison of web-mining methods

Suppose that a web-mining system evaluated a number of documents which were in the form of questionnaire, how would we realize that the system was accurate? We show the documents related to the questionnaire as relevant, and the retrieved document as the Retrieved. The collections of related and retrieved documents are shown as Relevant and Retrieved respectively.

There are 2 primary benchmarks to identify the quality of the retried text:

Accuracy: is the percentage of the retrieved documents related to the questionnaire, the benchmark is defined as follows:

$$recision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \quad (1)$$

Recall is the percentage of documents related to the questionnaire and actually retrieved. This benchmark is defined as follows:

$$Recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \quad (2)$$

6. Different types of Searching motors

The searching motor is a tool which makes it possible for us to search the information in the web. The searching motor usually offers a simple way for finding the information in the web, it doesn't search the World Wide Web direly it searches for a data base out of millions web pages.

The searching motors are divided into two primary groups: web crawls and web portal both have their own function swabbed on their specific features. The searching motors will be discussed and eventually their differences will be enumerated and eve will be aware of their uses.

6.1. Web Crawler

Since the web was created the web crawlers have been into being, the first crawler named Matthew was written in 1993. Web crawler starts working from some presupposed addresses and then receives their related pages from the web, in the next step, the links are detained from each page, and the links are then rearranged to be studied later. As you can see, all the words and phrases on a page will be indexed.

The data base can be searched later in order to answer the questions raised concerning the phrases on the page. Whenever the words and phrases on the page are saved in the data base. They will be searchable so every word or phrase can be searched on the internet. The data bases of crawlers are produced by spider. We call this operation "crawler" because all the links within a page are obtained for the next references. Although crawlers are believed to search the web to find the desired page, they actually remain in one place and locate the pages by following the links already situated on them.

Theoretically, the web crawler can start working from one particular page/ getting all the links out of it/ searching all links one after another until all pages on the net are examined. This strategy obviously runs into one difficulty. That is/ we can start out operation from one particular point and move to all anther parts and pages of the net a because some pages don't have any references and the

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.3,No.4, 2013
Published online: April 25, 2013
DOI: 10.7321/jscse.v3.n4.1

web crawler is not intelligent enough to realize which path to follow to search all the internet In fact, if there is no link from one page to another, spider reaches a deadened. The only way the web crawler can search a page is that the page is sent to it. The web crawlers usually work. Automatically and people don't usually keep them under control[13].

These features make the searching motors a suitable tool for finding particular information/ however, they are not quite efficient methods to obtain shared and common information.

6.2. web portal

There are different definitions for the web portal; there is no general agreement on one single definition. There is also dis-agreement on the name selected for it. Many believe that the word portal doesn't carry the meaning and it shouldn't be used, however, there is a general agreement on its particular features and operations.

It should be noted that the static web pages which provide links for the access to the heterogeneous sources related to a particular subject are actually directories and they shouldn't be mistaken for the web portrays. The web portals have got potentials and complexities in comparison with the directories.

In fact, the web portal doesn't include only one searching motor, it consists of different parts, providing many opportunities for the users, and the searching motor is one of them. In 2002, the Association of Research Libraries collected the data using a questionnaire sent to its members regarding the portals they created for their users[14, [1]].

The members stated that their desired portal would have two main features; first the portal should offer the users a searching motor that has the potential for searching different sources and collecting the results in one place and has at least one support system for the users. Based on the definition offered, only 19 members provided portals with the ability to search extensively and support. In the meantime a different research with a more specific definition of web portal was done, rendering 4 main operations and features. Out of 19 members, 16 members had the portals coming close to the above-mentioned definition. Based on the definition the web portals should involve the following four operations:

1) The potential for searching a unit in the web and Local database.
2) The efficient use of data crawling
3) The specific web page
4) The support services such as the management learning software and the ILI system.

Different universities rendered different descriptions of their portals; however, they are in agreement with each other on the operations performed at the web portal, their descriptions varied. It should be noted that there doesn't exist one general definition for the web portal and there exist some differences in the operations and services in some places.

Generally speaking, the web portal is defined as web database in which the information is classified based on the subject so that it will assist us with finding what we need: In other words, a web portal consists of a collection of related web pages and this collection is formed in a way that the information organization makes the access to the information and interaction with it possible.

Whenever one page is shown at the web portal, it is automatically collected and patterned so that they can be coordinated with the last content and the pattern chosen for that page. The great web portals try to supply all the classification related to one subject, but they don't always achieve that goal. The portals like the crawlers allow the possibility of their archives being searched, but the abstracts and the titles of web data bases are searched and the contents are not searched. The web portals are good for finding the general information. They are not able to organize everything in the web/so they can't help us obtain the specific information. This issue should be taken into consideration when the user decides to choose between a crawler and a web crawler[16].

The list of results is arranged in accordance with their probable priorities. The results are not always kept in order of their relativity and their quality. The expense a searching motor pays for them and other issues play a role in the ranking. A small number of the searching motors cover over 10% of

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.3,No.4, 2013
Published online: April 25, 2013
DOI: 10.7321/jscse.v3.n4.1

the web. None of them can cover more than 20%. The searching motors equipped with a given range and subjects have the potential to cover their subjects completely.

6.3. The comparison between the web portal's searching motors and web crawlers.

As it was mentioned earlier, the web portal and the web crawler each have the specific features and characteristics. We will compare the two mentioned groups of searching motor in this parts based on the table. The result is shown in table 3.

Evaluation Parameter	Web Portal	Web Crawler
The way of structuring	By human (by choosing pages)	By a computer program
The way of access	Subject index (the page are arranged based on the subject)	Without subject index (the pages are arranged by computer algorithm)
The indexing	Title, explanation, and abstract pages are indexed and searchable.	All pages are indexed and can be searched
Size	They are mostly small and rarely big	They are great and return a lot of information
The choice of information	The pages are often selected and provided after the initial evaluation	Information is varied and supplied without valuation
The possibility for searching	The possibility for simple searching with few operators	The possibility for complex searching with a lot of operators
The use	Obtaining the general information	Finding specific information on a single subject
Limitations	1) the insufficient coverage of web 2) the limited indexing	1) low accuracy of the information retrieved 2) limited coverage of the web 3) lack of result update.

7. Conclusion

We are faced with huge amount of data and database dive to the increase in the tecloniques and different wares in the production and collection of data, the importance of data base because of their accessibility and strength, and finally the World Wide Web as the main source of information.

There exist many problems on the path of efficient discovery of information on the web, for example the huge amount of information, the complexity of web pages in comparison with the common written document, the dynamic information on the web/ and establishing contact with million users worldwide. In spite of enormous amount of information on the web, small amount of it can be of use. How is it possible to measure the desired part of the web? How is it possible to obtain the high quality web page related to the particular subject? The above-mentioned problems have resulted in the studies concerning the discovery and the efficient use of the internet uses. Finally the solution has been rendered: web-crawling which can be performed by their main operations: searching the content searching the structure, and searching the use. A brief description of web crawling was offered in part 3.

In this article, the information on the choice of the operations of web Cawing. (SPAM), and the methods and techniques needed for the uses (T&A) were rendered. By using the tables and evaluation benchmark, an appropriate technique can be chased for the user in question. The procedure of information searching possible.

In fact, they supply the user with a simple solution to the process of finding information on the web. However, the searching motors face some deficiencies regarding the key words.

1) One title may involve hundreds or even thousands of documents; this problem causes it to return a lot of unrelated documents or low quality ones.

2) a lot of documents related to the subject in question may not contain the key words of that particular subject; this may lead to the problem of the meaning.

The above mentioned problems show that the present searching motors are sufficient to discover sources of information, but the use of different web crawling wares and methods plays an important role in improving the results of searching motor’s activities. Today many researchers conduct studies so as to enhance the function of these motors using the mentioned methods.

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.3,No.4, 2013
Published online: April 25, 2013
DOI: **10.7321/jscse.v3.n4.1**

References:

[1] M. Arroqui, C. Mateos, C. Machado, and A. Zunino, "RESTful Web Services improve the efficiency of data transfer of a whole-farm simulator accessed by Android smartphones," *Computers and Electronics in Agriculture*, vol. 87, pp. 14-18, 2012.

[2] A. Formica, "Semantic Web search based on rough sets and Fuzzy Formal Concept Analysis," *Knowledge-Based Systems*, vol. 26, pp. 40-47, 2012.

[3] B. Beldagli and T. Adiguzel, "Illustrating an ideal adaptive e-learning: A conceptual framework," *Procedia - Social and Behavioral Sciences*, vol. 2, pp. 5755-5761, 2010.

[4] A. ÇakIr, H. ÇalIs, and E. U. Küçüksille, "Data mining approach for supply unbalance detection in induction motor," *Expert Systems with Applications*, vol. 36, pp. 11808-11813, 2009.

[5] I. Colak, S. Sagiroglu, M. Demirtas, and M. Yesilbudak, "A data mining approach: Analyzing wind speed and insolation period data in Turkey for installations of wind and solar power plants," *Energy Conversion and Management*, vol. 65, pp. 185-197, 1// 2013.

[6] R. He, N. Xiong, L. T. Yang, and J. H. Park, "Using Multi-Modal Semantic Association Rules to fuse keywords and visual features automatically for Web image retrieval," *Information Fusion*, vol. 12, pp. 223-230, 2011.

[7] Y.-S. Hung, K.-L. B. Chen, C.-T. Yang, and G.-F. Deng, "Web usage mining for analysing elder self-care behavior patterns," *Expert Systems with Applications*, vol. 40, pp. 775-783, 2/1/ 2013.

[8] T.-T. Lee, C.-Y. Liu, Y.-H. Kuo, M. E. Mills, J.-G. Fong, and C. Hung, "Application of data mining to the identification of critical factors in patient falls using a web-based reporting system," *International Journal of Medical Informatics*, vol. 80, pp. 141-150, 2011.

[9] O. Nasraoui, C. Rojas, and C. Cardona, "A framework for mining evolving trends in Web data streams using dynamic learning and retrospective validation," *Computer Networks*, vol. 50, pp. 1488-1512, 2006.

[10] R. Olejnik, T.-F. Fortis, and B. Toursel, "Webservices oriented data mining in knowledge architecture," *Future Generation Computer Systems*, vol. 25, pp. 436-443, 2009.

[11] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Computers & Education*, vol. 51, pp. 368-384, 2008.

[12] Y.-H. Tao, T.-P. Hong, and Y.-M. Su, "Web usage mining with intentional browsing data," *Expert Systems with Applications*, vol. 34, pp. 1893-1904, 2008.

[13] Y.-H. Tao, T.-P. Hong, W.-Y. Lin, and W.-Y. Chiu, "A practical extension of web usage mining with intentional browsing data toward usage," *Expert Systems with Applications*, vol. 36, pp. 3937-3945, 2009.

[14] O. A. B. Penatti, E. Valle, and R. d. S. Torres, "Comparative study of global color and texture descriptors for web image retrieval," *Journal of Visual Communication and Image Representation*, vol. 23, pp. 359-380, 2012.

[15] Mehdi Bahrami, "Cloud Template, a Big Data Solution", The International Journal of Soft Computing and Software Engineering [JSCSE], Vol.3, No.2, pp.13-16. e-ISSN:2251-7545, DOI:10.7321/jscse.v3.n2.2

[16] T.-H. Wang, "Developing Web-based assessment strategies for facilitating junior high school students to perform self-regulated learning in an e-Learning environment," *Computers & Education*, vol. 57, pp. 1801-1812, 2011.

[17] Mehdi Bahrami, Amir Masoud Rahmani, Ahmad Faraahi, "*AGC4ISR, New Software Architecture for Autonomic Grid Computing*", 2010 IEEE International Conference on Intelligent Systems, Modelling and Simulation (ISMS), pp. 318-321,2010 doi:10.1109/ISMS.2010.65