ISSN:0975-9646

Building Domain Specific Search Engine Based on DLS and VSM Algorithms

Mohamed. M. Elbasyouni [1], Elsaeed. E. Abdelrazek[2], Doaa. M. Hawa[3]
Computer Teacher Preparation Department
Faculty of Specific Education, Damietta University, Egypt

***Abstract*— this paper aims to describe the proposed domain specific search engine based on Depth Limited Search (DLS) and Vector Space Model (VSM) algorithms specialize the content of crawled pages. It improved search result by filtering the web crawled pages and limit the depth of it using DLS algorithm and VSM algorithm. There are many difficult problems which are faced web search engines such as: spam, content quality, web conventions, and duplicate hosts. When doing research in Web, one typically uses search engines that are based on the Web crawling framework. Domain specific search engines use a special class of crawlers called "focused crawlers", for locating pages with a specific topic or related to a certain domain, the methodology in this paper suggests specific search engine using focused crawler depends on DLS and VSM algorithms ,This leads to information retrieval accurately to resolve the content quality problem.**

***Keywords*— Information retrieval, search tools, focused crawling, DLS, (Depth Limit Search), VSM (Vector Space Model)**

INTRODUCTION

Information retrieval (IR) is the task of representing, storing, organizing, and offering access to information items. IR is different from data retrieval, which is about finding precise data in databases with a given structure. In IR systems, the information is not structured; it is contained in free form in text (web pages or other documents) or in multimedia content. [1]

The web creates new challenges for information retrieval. The amount of information on the web is growing rapidly [2]; it contains information on many related and unrelated topics. [3]

There is a huge quantity of text, audio, video, and other documents available on the Internet, on about any subject. Users need to be able to find relevant information to satisfy their particular information needs. There are two ways of searching for information: search engines and directories organized by categories (such as Yahoo Directories). [1]

It has become increasingly difficult for users to find information on the WWW that satisfies their individual needs since information resources on the WWW continue to grow. Under these circumstances, Web search engines help users find useful information on the WWW. [4]

Web search engines collect data from the Web by "crawling" [18], specific search engines are based on focused crawlers, which collect only the documents related to the given topics of interest [19]

This paper describes the proposed domain specific search engine methodology based on Depth Limit Search (DLS) and Vector Space Model (VSM) algorithms specialize the content of crawled pages. It improved search result by filtering the web crawled pages and limit the depth of it using DLS algorithm and VSM algorithm.

II. PREVIOUS WORKS

[5] Develop the performance and quality of most current information retrieval (IR) system by handle three main axes: Quality of results by increasing relevant results (high precision), Usability and interactive presentation results, Performance includes response speed, cost, time and resources consuming.

[6] Introduce an intelligent Adaptive focused crawling strategy. The proposed crawler is intelligent as it can estimate the relevancy of a web page before actually visiting it. It is also adaptive as it keeps track with any changes that may arise in its domain of interest.

[7] Develop a latent semantic indexing classifier that combines link analysis with text content in order to retrieve and index domain specific web documents. And combined links and terms in an LSI based algorithm. According to the study LSI based algorithm depends on the size of trained document data and it's not recommended to start with small size list.

[8] Introduces a simple focused crawler, which is described by two parameters, degree of relatedness, and depth. Both provide an opportunity for the crawler to "tunnel" through lowly ranked pages. This study also describe the two types of web crawling strategies deployed by web search engines, breadth first search strategy and best first search strategy. Breadth first search strategy endeavours to build a general index of the web covering any conceivable topic by endeavouring to search a significant portion of the web. The best first search focuses on the retrieval of pages which are relevant to a particular given topic. A crawler using a best first search strategy is known as a "focused crawler".

[9] Introduced a simple framework for focused crawling using combination of two existing methods, the Link Structure analysis and Content Similarity. Our generic framework is more powerful and flexible than previously known focused crawlers.

[10] Describes the basic task performed search engine. Overview of how the Web crawlers are related with search engine. And introduce a basic types of search engine such as Crawler Based Search Engines , Human Powered Directories and Hybrid Search Engine Also introduce crawling techniques for example Focused Crawling and Distributed Crawling .

[11] Supporting the researcher to access relevant and interesting information, the presented system proposes five types of intelligent agents: User Interface Agent, Intelligent

Recommendation Agent, Knowledge Agent, Search Agent and Filter Agent. The filtering process uses the back propagation technique of errors to adjust the connection weights of a neural network with three layers. The elements of the neural network input vector are simply the occurrences of terms in articles .

[12] Introduces the recent advances in the information retrieval systems. In addition, a Term Weight information retrieval model is proposed. The proposed model, Term Weight information retrieval model based on Document Structure (TW-DS), provides a significant enhancement on the most common information retrieval model which is the vector space model. The proposed model is enhanced with a new term weighting strategy which is based on the structure of the document. The major advantage of the proposed model is the treatment of the document as separated parts in such a way that the importance of terms in each part is determined by a term frequency.

III. THE PROPOSED SYSTEM

The proposed system was called Search Engine for Artificial Intelligence Field (SEAIF). It improved search result by filtering the web crawled pages and limit the depth of it using depth limited search (DLS) algorithm and Vector Space Model (VSM) algorithm. Depth Limited Search (DLS) is an algorithm to explore the vertices of a graph, it is an uninformed search. Vector Space Model (VSM) algorithm used in information filtering, information retrieval, indexing and relevancy rankings.

The proposed system consists of five phases, and each phase contains some steps.

1. First Phase: Optimum Depth Calculation

The regular search engine used to crawl the web pages discarding the categorization and quality of the pages, this phase uses depth limited search algorithm to determine the best depth required to insure that the crawled pages are highly related to the category specified.

The first phase is illustrated in Figure 1

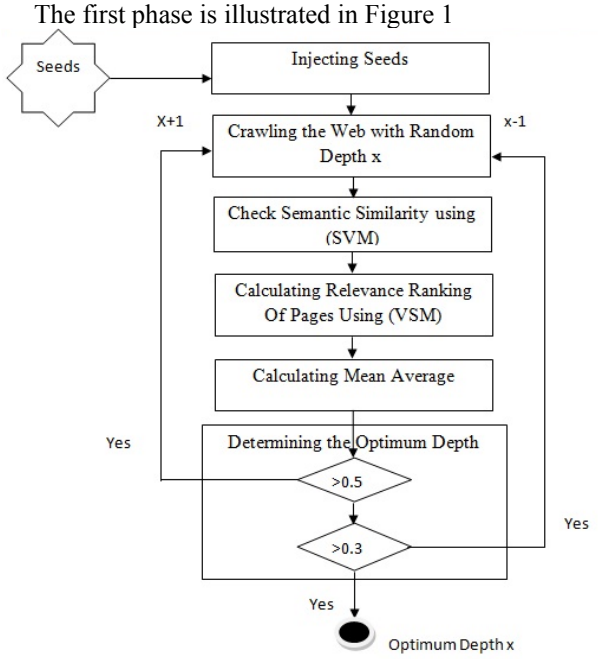

Fig. 1: Optimum depth calculation phase

Optimum depth calculation phase aims to determine the optimum depth by (6) steps, which will be discussed in the following part.

1- *Injecting Seeds*

The system administration has to collect some seeds that are related to the category to start the crawling with.

2- *Crawling the Web with Random Depth*

The system crawl the seeds with random depth and indexing data result into index.

3- *Check Semantic Similarity*

By using SVM algorithm (support vector machine) given a set of training pages marked as belonging to one of computer science categories, an SVM training algorithm builds a model that assign new pages into one category or another.

An SVM classifier attempts to maximize the following function with respect to $\vec{w}$ and b: [13]

$$Lp = \frac{1}{2}\left\|\vec{w}\right\| - \sum_{i=1}^{t} \alpha_i y_i \left(\vec{w} \cdot \vec{x}_i + b \right) + \sum_{i=1}^{t} \alpha_i$$

where t is the number of training examples, and α_i, i = 1, …., t, and non-negative numbers such that the derivatives of Lp with respect to α_i are zero; α_i are the Lagrange multipliers and Lp is called the Lagrangian. In this equation, the vectors $\vec{w}$ and constant b define the boundary.

4- *Calculating Relevance Ranking of Pages Using(VSM)* After checking the semantic similarity the pages tend to calculate the cosine of the angle between pages represented as vectors instead of the angle itself, documents and queries are represented as vectors.

The term weighting " w_i " can be given by the following equation: it is called "term frequency-inverse document frequency "(tf- idf) [14]

$$w_i = tf_i \times idf_i$$

Where,

tfi : ith term frequency

idfi : inverse document frequency of ith term

And

$$idf_i = \log \left(\frac{D}{df_i} \right)$$

Where

log : logarithm.

D : Number of documents.

df_i : Number of documents containing term i.

The degree of similarity between the documents and queries is defined as cosine similarity (cos (θ)) and can be calculated by the following formula: [15]

$$sim(d,q) = \cos(\theta) = \frac{d \bullet q}{\|d\|\|q\|} = \frac{\sum_{i=1}^{n} d_i \times q_i}{\sqrt{\sum_{i=1}^{n} (d_i)^2} \times \sqrt{\sum_{i=1}^{n} (q_i)^2}}$$

Where,

Θ : Angle between two vectors d & q.

d : A document.
q : A query.

5- *Calculating Mean Average*

After calculating the relevance ranking of each page using VSM the proposed system calculates mean average of all the pages crawled using X depth.

6- *Determining The Optimum Depth*

After calculating the mean average the proposed system result in three different options :

The mean average is higher than 0.5, repeat steps 2:5 using depth X+1

The mean average is below 0.3, repeat steps 2:5 using depth X-1

The mean average is below 0.3 and higher than0.5 return to the depth X as the optimum depth limit

2. second phase: crawling the web

This phase aim to crawl the web using optimum depth limit which calculated from phase one, this phase consist of (4) steps as follows: Injecting seeds - Fetching - Parsing – Analyzing

The second phase is illustrated in Figure 2

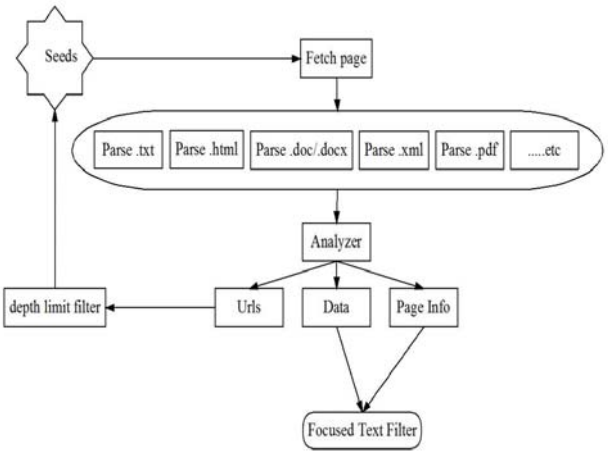

Fig. 2: Crawling the web phase

The crawler begins with the URL that constitute a seed set it picks a URL from this seed set then fitches the web at that URL , the fetched page is then parsed to extract , both the text and the links from the page , The links from the page are added to the seeds once again if and only if it satisfy the condition (lower than the optimum depth), and that links will be crawled again in continuous way, The crawled data from the page are then added to the index alongside with the page information itself and the location in the page where the data were found, The recursive process of crawling web pages will continue until no more new URLs in the seed set or disk full and that require to spend a lot of time waiting for responses to requests and this process is so complicated by the many demands on a practical web crawling system, Another issue is that the crawling systems will waste the resources of different web servers.

In order to satisfy the need of efficient, scalable, distributed, polite, the proposed system uses map reduce algorithm, to use threads and fetch hundreds of pages at once.

The Map and Reduce are both defined with respect to data structured in (key, value) pairs [16]

$$\text{Map } (k1, v1) \rightarrow \text{list } (K2, v2)$$

$$\text{Reduce } (K2, \text{list } (v2)) \rightarrow \text{list } (v3)$$

The basic idea of the map reduce technique is to send a job to master node , the master breaks the work up into pieces and sends it (mapping) to the various worker nodes (other threads) which perform there assigned sub-tasks , and then sends the sub-result back to master once the master node gets all the sub-result it starts to combine them (reducing) in to final result, all of these tasks are done at the same time and each computer is given the right amount of work to keep it occupied the whole time.

The result of this phase will be injected in the next phase and it consists of three results:

1-Page data: The real content of the page.

2- Page info: URL of the page itself and header information and domain information.

3-URLs located within each page: All the hyper links located in each page.

Result one and two are injected into phase three to filter the content of pages, Result three are compared to the optimum depth limit and injected in the seeds if and only if the depth of URL is below / lower the optimum depth limit

3. THIRD PHASE: FOCUSED FILTERING PHASE

This phase aims to filtering the content of the crawled data to insure that the data is related to the specific category.

The third phase is illustrated in Figure 3

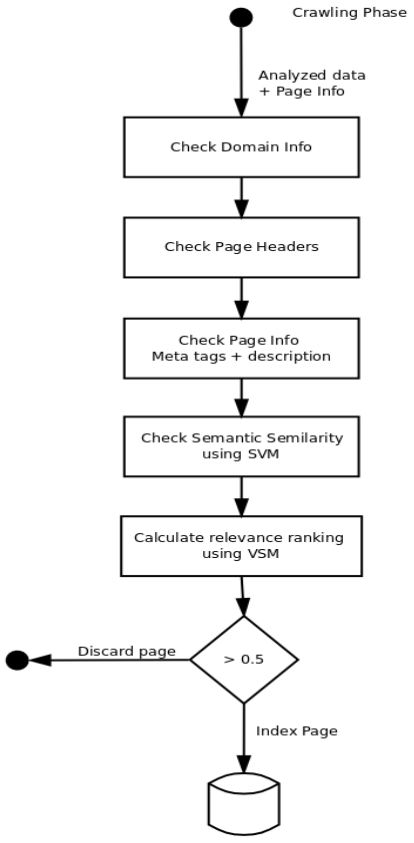

Fig. 3: Focused filtering phase

This phase consists of (5) different steps which will be discussed as follow:

1-Check Domain Information

The domain is the name of a network associated with an organization which points to the corresponding IP address of the organization server.

2-Check Page Headers

Computers send an http request to the server and the server response with an http response, these http requests and responses holds some information about the server IP and the page required from that server contained in http fields. In this step the proposed system will check each crawled page to insure accuracy of the data gathered from that page by checking the http request header and http response header for full data a transferred and connection closing

3-Check Page Information

In this step check the Meta tags, keywords, descriptions and the headings of the page to insure the crawled data is related to the artificial intelligence category.

4-Check Semantic Similarity Using (SVM) algorithm

Using the word net as the training data of the SVM algorithm, The SVM algorithm builds a model that assign new data to the specific category or another and then storing the data into indexes.

5-Calculate Relevance Ranking Using VSM (Vector Space Model)

In this step calculate the relevance ranking between the data and the specific category using VSM algorithm , then indexing the data if and only if the ranking is higher than 0.5 otherwise it will discard it.

4. FORTH PHASE: INDEXING

This phase aims to collect and store data to facilitate fast and accurate information retrieval using inverted index algorithm and uses the full text indexing to store a list of references to documents for each word and the position each word with in a document.

5. FIFTH PHASE: SEARCHING THE WEB

This phase aims to search into indexed pages and match the query string keywords that were entered by the user, with the indexed data.

The fifth phase is illustrated in Figure 6

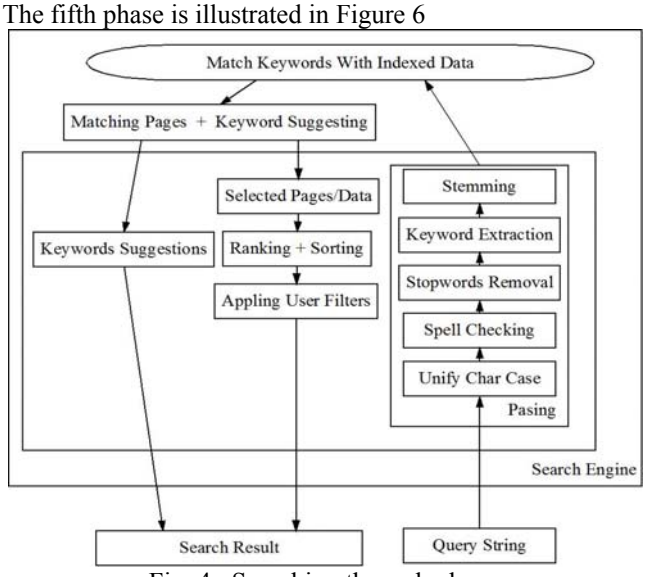

Fig. 4: Searching the web phase

This phase consists of two main sub-phases each of them contain more extra steps

First sub-phase

1-UNIFY CHAR CASE

Different char case may result in different matched pages or no matching at all, to solve this issue transform the query string char case to lower case as the indexed data.

2-SPELLCHECKING

One of the better ways to enhance the search experiences is by offering spelling correction.

3-STOP WARDS

Common words such as " a , an , and , ….etc " add little value to a search result but since these words are so common , removing them will contribute considerably to the searching time and searching result .

4-KEYWORDS EXTRACTION

Keywords are the important topics in the query string and can be used to index data, generate tag clouds or for searching keywords extraction algorithm employs sophisticated statistical algorithm and natural language processing technology to analyze the data and automatically extract relevant terms from a given corpus.

5-STEMMING

The stemming step is based on removing the suffixes which are mostly made up of a combination of smaller and simpler part which done by some basic rules.

The matching of the keywords with indexed data

The matching step scan the stemmed keywords of the query string then generate a list of matching pages from the index then return the matched pages to the second sub-phase to be processed again then send back to the user .

THE SECOND SUB-PHASE

1- *Ranking*

Ranking by page rank algorithm, Page rank is one of the method used to determine a page relevant or importance it assigns a numerical weighting to each page in this step the (AISE) system calculate the page rank of each matched result and assign that number to the page in order to be used in the sorting step, Ranking the retrieved pages plays an important rule in the sorting step.

The Page Rank of a page A is given as follows: [17]

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Where:

PR (A) is the Rank of page A,

PR (T1) is the Page Rank of pages Ti which link to page A,

C(T1) is the number of outbound links on page Ti and d is a damping factor which can be set between 0 and

2- Sorting

The sorting is required to display the retrieved pages in proper order so that the most relevant and accurate pages displays first.

The sorting is required to display the retrieved pages in proper order so that the most relevant and accurate pages displays first.

3- Filtering

on the last step before sending the result back to the user is applying the user filter , user filters may change the sorting of the retrieved pages according to date , place or a file.

IV. EXPERIMENTAL RESULTS

The experiment was applied in the "Computer Laboratory, Faculty of Specific Education, Damietta University" The graphical user interface of the proposed system is shown in Figure 5 This figure shows the text box and some buttons such as (Home, About, Search). The user enters the sentence in the text box. After s (he) checks the button "Search". Then the results of the sites are displayed in the user interface ".

Sample of searching results are shown in figure 6

Fig.5: The Graphical user interface of the proposed system

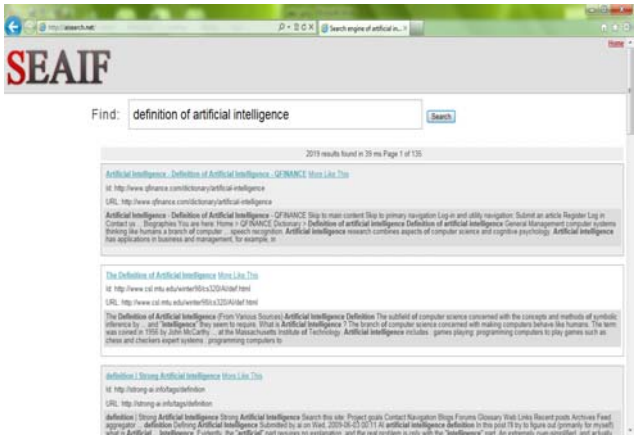

Fig.6: Sample of searching results sites

SYSTEM PERFORMANCE EVOLUTION

The system was tested by twenty specialists in computer science. Statistical processing based on (A large degree of correlation, Medium degree of correlation, weak degree of correlation between query and results). The comparing between the Proposed system and Google based on the first twenty sites from each.

Sample of results is shown in table1

TABLE I

SAMPLE RESULTS OF MEASUREMENT

No. Of Key Words	Google								Proposed System (SEAIF)							
	Accuracy (Large Degree)		Accuracy (Medium Degree)		Accuracy (Weak Degree)		Chi2	Asymp Sig	Accuracy (Large Degree)		Accuracy (Medium Degree)		Accuracy (Weak Degree)		Chi2	Asymp Sig
	F	%	F	%	F	%			F	%	F	%	F	%		
1	245	61.3	104	26.0	51	12.8	150.81	.000	323	80.8	48	12.0	29	7.3	406.05	.000
2	288	72.0	66	16.5	46	11.5	270.62	.000	359	89.8	36	9.0	5	1.3	576.51	.000
3	264	66.0	79	19.8	57	14.3	193.89	.000	331	82.8	43	10.8	26	6.5	440.64	.000
4	244	61.0	96	24.0	60	15.0	142.64	.000	307	76.8	72	18.0	21	5.3	349.05	.000
5	321	80.3	54	13.5	25	6.3	399.36	.000	351	87.8	46	11.5	3	.8	539.94	.000
……..	….	….	….	….	….	….		….	….	….	….	….	….	….		….

TABLE2

Sample Results Of Sign TestNo. Of Key Words	Negative Differences [a]	Positive Differences [b]	Ties [c]	Total
1	0	100	300	400
2	0	107	293	400
3	0	98	302	400
4	0	102	298	400

a SEAIF < Google b SEAIF > Google c SEAIF = Google

CONCLUSION

This paper explains the proposed domain specific search engine methodology based on Depth Limited Search (DLS) and Vector Space Model (VSM) algorithms specialize the content of crawled pages. it uses a special class of crawlers called "focused crawlers", for locating pages with a specific topic or related to a certain domain depends on DLS and VSM algorithms as a try to resolve the content quality problem, introduce sample of results measurement and sample results of sign test.

REFERENCES

[1] D. Inkpen : " Information Retrieval on the Internet" , Ph.D. ,University of Toronto ,Canada , 2006. available: http://www.site.uottawa.ca/~diana/csi4107/IR_draft.pdf

[2] S. Brin and L. Page : " The Anatomy of a Large-Scale Hypertextual Web Search Engine", computer network and ISDN systems, Vol.30, No. 1-7 Stanford, USA, 2006.

[3] P.M.E. De bra and R.D.J. Post: "Information Retrieval in the World−Wide Web: Making Client−based searching feasible", computer network and ISDN systems, Vol.27, No.2, 2004.

[4] K. Sugiyama, K. Hatano and M. Yoshikawa : " Adaptive Web Search Based on User Profile Constructed without Any Effort from

Users", Takayama, Ikoma, Japan conference on World Wide Web, 2004

[5] A. A. R. Slamaa: "Information visualization for information retrieval based on fast search technique", Thesis (M. S.), Information Systems Department, Faculty of Computers and information science, Mansoura University,Egypt, 2013.

[6] A. I.M.Saleh: "Building of a domain specific search engine", Thesis (Ph.D.) ,Department of Computers Engineering and Systems , Faculty of Engineering , Mansoura University,Egypt,2006.

[7] G. Almpanidis and C. Kotropoulos "Combining Text and Link Analysis for Focused Crawling", Information SystemsVol. 32, No. 6, September 2007.

[8] A.C.Tsoi, D. Forsali, M. Gori, M. Hagenbuchner and F. Scarselli : "A Simple Focused Crawler", Universita' degli studi di Siena Siena, Italy,2003. Available:http://cs.brynmawr.edu/Courses/cs380/fall2006/wwww12 DanielePoster.pdf

[9] M. Jamali, H. Sayyadi, B. B. Hariri and H. Abolhassani:" A Method for Focused Crawling Using Combination of Link Structure and Content Similarity", ieeexplore.ieee.org, Web Intelligence, IEEE/WIC/ACM International Conference on 2006.

[10] M.P.S.Bhatia and D.Gupta:"Discussion on Web Crawlers of Search Engine", Proceedings of 2nd National Conference on Challenges & Opportunities in Information Technology (COIT-2008) RIMT-IET, Mandi Gobindgarh. March 29, 2008.

[11] H. H. A. El-Hadidi:" Intelligent Agent System For Navigation Assistance And Information Retrieval", Thesis (Ph.D.), Department Of Mathematics, Faculty Of Science In Damietta, Mansoura University,Egypt, 2009.

[12] S. M. Ahmed:" Web Based Information Retrieval Mode", Thesis (Master), Department Of Electronics, Communications & Computers, Faculty Of Engineering, Helwan University,Egypt,2009.

[13] X.Wu ,V. Kumar ,J. R. Quinlan , J. Ghosh , Q. Yang , H. Motoda , G. J .McLachlan ,A. Ng , B. Liu , P. S.Yu , Z.Zhou , M. Steinbach , D. J. Hand and D. Steinberg : " Top 10 algorithms in data mining", Knowl Inf Syst , vol. 1 .No. 14,2008.

[14] A. A . Ewees :"A Proposed Expert System For Assessing Students Arabic Essay Using Data Mining", Thesis (Ph.D.), Computer Teacher Preparation Department , Faculity Of Specific Education , Mansoura University ,Egypt,2012.

[15] T.K. Landauer,D..Laham.andP.Foltz,,"Automatic Essay Assessment" ,Assessment in Education,Vol.10,No.3,2003.

[16] J. Dean and S. Ghemawat:" MapReduce: Simplified Data Processing on Large Clusters", published in magazine Communications of the ACM, Vol. 51 No. 1, January 2008, ACM New York.

[17] D.Akshata ,R.. Deore and L. Paikrao : " Ranking Based Web Search Algorithms", International Journal Of Scientific And Research Publications, Vol. 2, No. 10, October 2012.

[18] B. Novak:" A Survey Of Focused Web Crawling Algorithms",In: SIKDD 2004 at multiconference IS 2004, 12-15 Oct 2004, Ljubljana, Slovenia. Available: http://eprints.pascalnetwork.org/archive/00000738/01/BlazNovak-FocusedCrawling.pdf

[19] A. Micarelli and F. Gasparetti:" Adaptive Focused Crawling", Lecture Notes in Computer Science Vol. 4321, 2007. Available : http://link.springer.com/chapter/10.1007%2F978-3-540-72079-9_7#