ISSN (Online): 2455-7838
SJIF Impact Factor (2016): 4.144

EPRA International Journal of

Research & Development

(IJRD)

Monthly Peer Reviewed & Indexed International Online Journal

Volume:2, Issue:4, April 2017

Published By :
EPRA Journals

SJIF Impact Factor: 4.144 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 2 | Issue: 4 | April | 2017

PROPOSED SYSTEM FOR INFORMATION RETRIEVAL FROM THE INTERNET
(Accurately, quickly and easily)

Doaa. M. Hawa[1]
[1] Doctor of Computer Teacher Preparation Department, Faculty of Specific Education, Damietta University, Egypt

Abeer. M. Saad [2]
[2] Doctor of Computer Teacher Preparation Department, Faculty of Specific Education, Damietta University, Egypt

ABSTRACT

This paper aims to provide proposed system for searching the web in artificial intelligence field to access the required information accurately, quickly and easily.

KEYWORDS: *Information retrieval, search tools, crawling, searching.*

I. INTRODUCTION

Retrieving information from the web is becoming a common practice for internet users[3], the huge size and heterogeneity of the web is no longer in doubt, Therefore, the web poses a dire challenge to the effectiveness of classical information retrieval systems[2], A critical goal of successful information retrieval on the web, though, is to identify which pages are of high quality and relevance to a user's query, The success of the web lies in the many software tools that are available for its information retrieval, These software include the search engines (Google, AltaVista etc), hierarchical directories (Yahoo), many other software agents and collaborative filtering systems [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" [8], specific search engines are based on focused crawlers, which collect only the documents related to the given topics of interest [9]

This paper provides proposed system for searching the web in artificial intelligence field to access the required information accurately, quickly and easily.

II. PREVIOUS WORKS

This thesis aims to develop the performance and quality of most current 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. The thesis presents solution for the following problems: Web search results usually have low precision (Quality of retrieved information), All users are not created equal (Different users may use different terms to describe similar information needs), Query formulations, Information overload, Quality performance of retrieving, and Presentation formats of the retrieval results. The thesis proposes using conceptual model applied in two phases: first phase, query expansion using Word Net Ontology to enhance and rich user's query (one term or more) also dealing with meaning. Second phase, is the concepts in each document. Query concepts are matched with the document concepts in final phase. The challenge is reach to what user means not what he writes, so user gets more relevant results than previous. In another words, the system is high precision, it equals to high quality. Information visualization is a key solution for poor interface (presentation results) which practical application of it in computer programs involves selection, transforming, and representing abstract data in a form that facilitates human interaction for exploring and understanding. Similarity, using it in IR on the web will be helpful and contribute to increase user's satisfaction. Regional crawler in distributed environment and personalization support achieve high performance. One side of that, it made system present offline search (results) at first if user's query similar to another previously answered, else fire online search. Also instead of one crawler and one index, there are many crawlers work in parallel distributed regionally. from that, system gets high response time, low cost, low time and resources consuming, then high performance[6].

[5] 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.

This study presents an overview of the focused crawling domain and, in particular, of the approaches that include a sort of adaptively. That feature makes it possible to change the system behavior according to the particular environment and its relationships with the given input parameters during the search. Particular attention has been given to adaptive focused crawlers, where learning methods are able to adapt the system behavior to a particular environment and input parameters during the search. Evaluation results show how the whole searching process may benefit from those techniques, enhancing the crawling performance. Adaptively is a must if search systems are to be personalized according to user needs, in particular if such needs change during the human-computer interaction. Besides new crawling strategies that take into consideration peculiar characteristics of the Web, such as topical locality, we can expect future research to head in several directions. Some techniques used to look into the dark matter (hidden Web), and Natural Language Processing (NLP) analysis can help understand the content of Web pages and identify user needs. In this way, the effectiveness of crawlers can be improved both in terms of precision and recall[7].

This study 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].

This study describe the use of ontology-supported Web site models to provide a semantic level solution for a search agent so that it can provide fast, precise, and stable search results. It features the following interesting characteristics, including ontology-supported construction of website models, website models supported Website model expansion, and website models-supported Webpage Retrieval. In addition, our ontology construction is based on a set of pre-collected web pages on a specific domain; it is hard to evaluate how critical this collection process is to the nature of different domains. We are planning to employ the technique of automatic ontology evolution to help studying the robustness of our ontology[9].

III. THE PROPOSED SYSTEM

The proposed system consists of five phases:

First phase : optimum depth calculation
Second phase : crawling the web
Third phase : focused filtering
Forth phase : indexing
Fifth phase : search the web

The proposed system was called search engine for Artificial Intelligent field (SEAIF) system.
It improved search result by filtering the web crawled pages and limit the depth of it dependent on the merger between depth limit search (DLS) algorithm and Vector Space Model (VSM).

THE PROPOSED SYSTEM OVERVIEW

The proposed system contains five phases; these phases are illustrated in Figure 1

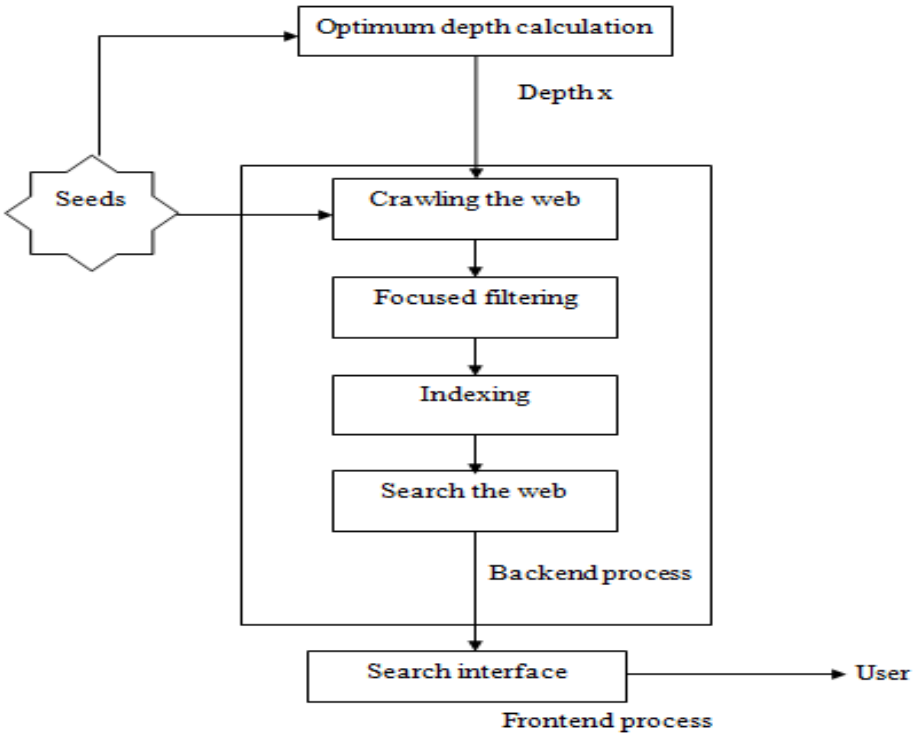

Figure 1: The proposed system phases (SEAIF)

FIRST PHASE: OPTIMUM DEPTH CALCULATION

This phase aims to determine the best depth required to insure that the crawled pages are highly related to the category specified as (Artificial intelligent) by using depth limited search algorithm.

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 four steps as follows

- Injecting seeds
- Fetching
- Parsing
- Analyzing

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 category of artificial intelligence.

FORTH PHASE: INDEXING

This phase aims to collect and store data to facilitate fast and accurate information retrieval using inverted index. The system uses the full text indexing to store a list of document references. Thos references refer to each word in the document and the position of each word within that document.

FIFTH PHASE: SEARCHING THE WEB

This phase aims to search imperviously indexed pages and match the query string keywords that were entered by the user, with the indexed data. During this phase the (SEAIF) system process the query string before matching that query string with indexed data.

IV. EXPERIMENT

The experiment tested SEAIF system efficiency then compared the result with Google search engine. The procedures are described below:

- Twenty phrase has been searched for in the Google search engine and search for them again in the proposed system by a sample of faculty members and assistants, Department of teacher preparation Computer University Faculty of Specific Education Damietta.
- In each sentence was tested first 20 of the total output of the search results for this phrase in the Google search engine and also the first 20 of the total output of the search results for this sentence in the proposed system.
- Every member of the research sample was conducted by a comparison between the Google search engine and the proposed system in terms of the total number of results, speed (response time) and accuracy of search results retrieved for each of them and each phrase separately.
- All the result of the search results was examined by three responses Are (largely accurate - moderately accurate - accurate weakly)

Table 1: results before being processed statistically (accuracy)

N	Keywords	SEAIF proposed system			Google search engine		
		accuracy			accuracy		
		Accuracy (weak degree)	Accuracy (medium degree)	Accuracy (large degree)	Accuracy (weak degree)	Accuracy (medium degree)	Accuracy (large degree)
1	Definition of artificial intelligence	29	48	323	51	104	245
2	Applications of artificial intelligence	5	36	359	46	66	288
3	History of artificial intelligence	26	43	331	57	79	264
4	Development of artificial intelligence	21	72	307	60	96	244
5	Language of artificial intelligence	3	46	351	25	54	321
6	Artificial intelligence in education	1	27	372	11	52	337
7	Definition of expert system	23	35	342	47	71	282
8	Expert system in education	7	72	321	40	95	265
9	Image processing	0	29	371	22	38	340
10	What is genetic programming	0	3	397	2	38	360
….	……..	….	….	…..	…..	…..	…..

Table 2: results before being processed statistically (number)

N	Keywords	Number results in (SEAIF) proposed system	Number results in (Google)search engine
1	**Definition of artificial intelligence**	2019	22700000
2	**Applications of artificial intelligence**	6038	44000000
3	**History of artificial intelligence**	5397	44300000
4	**Development of artificial intelligence**	5506	76600000
5	**Language of artificial intelligence**	5760	70700000
6	**Artificial intelligence in education**	4388	84200000
7	**Definition of expert system**	1460	30400000
8	**Expert system in education**	1665	74600000
9	**Image processing**	4527	221000000
…	……..	……	……..

Table 3: results before being processed statistically (response time)

N	Key words	response time in (SEAIF) proposed system	response time in (Google)search engine
1	**Definition of artificial intelligence**	28.2 ms	362.5 ms
2	**Applications of artificial intelligence**	17.35 ms	341.5 ms
3	**History of artificial intelligence**	32 ms	399 ms
4	**Development of artificial intelligence**	31.1 ms	283 ms
5	**Language of artificial intelligence**	28.35 ms	369 ms
6	**Artificial intelligence in education**	28.25 ms	386.5 ms
7	**Definition of expert system**	28.8 ms	397 ms
8	**Expert system in education**	28.45 ms	369 ms
9	**Image processing**	16.75 ms	263 ms
…	……	……..	……..

RESULTS

The results showed that there is a statistically significant difference at the 0.05 level of significance between the Google search engine and the proposed system (SEAIF) regarding the total number of results, speed (response time) and accuracy in favor of the proposed system.

There is an increase in speed in the proposed system (SEAIF) for the Google search engine.

The number of results in the proposed system (SEAIF) is less than the Google search engine.

And there are not any results in less accuracy in the proposed system (SEAIF) for the Google search engine, the percentage of accuracy in the proposed system is 90.24% and the percentage of accuracy in the Google search engine is 75.76%, this indicates that the search results in the proposed system with higher accuracy than Google search engine.

Table 4: accuracy of results between the Google search engine and the proposed system

No. of Key words	Google								Proposed system (SEAIF)								correlation
	Accuracy (large degree)		Accuracy (medium degree)		Accuracy (weak degree)		Chi^2	asymp sig	Accuracy (large degree)		Accuracy (medium degree)		Accuracy (weak degree)		Chi^2	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	.793
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	.757
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	.810
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	.811
5	321	80.3	54	13.5	25	6.3	399.36	.000	351	87.8	46	11.5	3	.8	539.94	.000	.824
6	337	84.3	52	13.0	11	2.8	472.95	.000	372	93.0	27	6.8	1	.3	643.35	.000	.731
7	282	70.5	71	17.8	47	11.8	250.80	.000	342	85.5	35	8.8	23	5.8	490.38	.000	.809
8	265	66.3	95	23.8	40	10.0	206.37	.000	321	80.3	72	18.0	7	1.8	412.05	.000	.791
9	340	85.0	38	9.5	22	5.5	481.46	.000	371	92.8	29	7.3	0	.0	638.61	.000	.831
..	….	….	…	…..	…..	…..	…….	…..	….	…..	…..	…..	….	….	……	….	…..

Table 5: Sign test between the Google search engine and the proposed program

No. of 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
5	0	45	355	400
6	0	52	348	400
7	0	89	311	400
8	0	84	316	400
9	0	53	347	400
....	.....	.....	....	...

Table 6: Speed in the Google search engine and the proposed system (SEAIF)

		N
google_time - seaif_time	Negative Differences[a]	0
	Positive Differences[b]	20
	Ties[c]	0
	Total	20

a. google_time < seaif_time

b. google_time > seaif_time

c. google_time = seaif_time

Table 7: numbers in Google search engine and the proposed system (SEAIF)

		N
google_numbers - seaif_numbers	Negative Differences[a]	0
	Positive Differences[b]	20
	Ties[c]	0
	Total	20

a. google_numbers < seaif_numbers

b. google_numbers > seaif_numbers

c. google_numbers = seaif_numbers

IIV. CONCLUSION

This paper describes the proposed system for searching the web in artificial intelligence field the proposed system called Search Engine for Artificial Intelligence Field (SEAIF), the proposed system specialize the content of crawled pages. It improved search result by filtering the web crawled pages and limit the depth of it dependent on merging between DLS algorithm and VSM algorithm as a try to resolve the content quality problem, introduce sample of results measurement and sample results of sign test

It consists of five phases, and each phase contains some steps First phase optimum depth calculation, Second phase crawling the web, Third phase focused filtering , Forth phase indexing, Fifth phase search the web.

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. 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.*
6. *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/wwww12DanielePoster.pdf*
7. *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.*
8. *B. Novak:" A Survey Of Focused Web Crawling Algorithms",*
9. *In: SIKDD 2004 at multiconference IS 2004, 12-15 Oct 2004, Ljubljana, Slovenia. Available:*
10. *http://eprints.pascalnetwork.org/archive/00000738/01/BlazNovak-FocusedCrawling.pdf*
11. *A. Micarelli and F. Gasparetti:" Adaptive Focused Crawling", Lecture Notes in Computer Science Vol. 4321, 2007. Available :*
12. *http://link.springer.com/chapter/10.1007%2F978-3-540-72079-9_7#*