META SEARCH ENGINE WITH AN INTELLIGENT INTERFACE FOR INFORMATION RETRIEVAL ON MULTIPLE DOMAINS

D.Minnie[1], S.Srinivasan[2]

[1]Department of Computer Science, Madras Christian College, Chennai, India
minniearul@yahoo.com
[2]Department of Computer Science and Engineering, Anna University of Technology Madurai, Madurai, India
sriniss@yahoo.com

ABSTRACT

This paper analyses the features of Web Search Engines, Vertical Search Engines, Meta Search Engines, and proposes a Meta Search Engine for searching and retrieving documents on Multiple Domains in the World Wide Web (WWW). A web search engine searches for information in WWW. A Vertical Search provides the user with results for queries on that domain. Meta Search Engines send the user's search queries to various search engines and combine the search results. This paper introduces intelligent user interfaces for selecting domain, category and search engines for the proposed Multi-Domain Meta Search Engine. An intelligent User Interface is also designed to get the user query and to send it to appropriate search engines. Few algorithms are designed to combine results from various search engines and also to display the results.

KEYWORDS

Information Retrieval, Web Search Engine, Vertical Search Engine, Meta Search Engine.

1. INTRODUCTION AND RELATED WORK

WWW is a huge repository of information. The complexity of accessing the web data has increased tremendously over the years. There is a need for efficient searching techniques to extract appropriate information from the web, as the users require correct and complex information from the web.

A Web Search Engine is a search engine designed to search WWW for information about given search query and returns links to various documents in which the search query's key words are found. The types of search engines [1 – 3] used in this paper are General Purpose Search Engines such as Google, Vertical Search Engines and Meta Search Engines.

A Vertical Search Engine searches the web and returns results for specific queries on a domain. The filter component of Vertical Search Engine classifies the web pages downloaded by the crawler into appropriate domains [4]. Medical search engines such as "MedSearch", facilitate the user to get information on medical domain [5 - 7]. Vertical Search Engines such as iMed creates search queries for the user by interacting with the user [8].

The various Search Engines follow different page ranking techniques and hence the user may fail

DOI : 10.5121/ijcseit.2011.1404

to receive the appropriate information. This gives rise to the use of Meta Search Engines [9]. Meta Search Engine accepts a search query from the user and sends the search query to a limited set of Search Engines. The results retrieved from the various search engines are then combined to produce a result set.

Information Retrieval is the science of searching and retrieving information from documents. Web search is also a type of information retrieval as the user searches the web for information. The efficiency of a search facility is measured using two metrics Precision and Recall. Precision specifies whether the documents retrieved are relevant and Recall specifies whether all the relevant documents are retrieved.

$$\text{Precision} = \text{Relevant and Retrieved} / \text{Retrieved} \quad (1)$$

$$\text{Recall} = \text{Relevant and Retrieved} / \text{Relevant} \quad (2)$$

Precision is calculated as the ratio of the relevant web pages that are retrieved to the number of web pages that are retrieved. Recall is not calculated as it is not possible to find the number of relevant web pages out of the millions of web pages in the WWW.

2. WEB SEARCH ENGINE (WSE)

A Web Search Engine consists of a Web Crawler, Indexer, Query Parser and Ranker and its architecture is shown in Fig. 1. WSE stores information about many web pages, which they retrieve from the WWW itself. These pages are retrieved by a Web Crawler which follows every link it sees. The contents of each page are then analyzed to determine how it should be indexed. The crawling and indexing operations are performed at back-end. At the front-end, the user is allowed to search the web for a specific query. The query is parsed and an appropriate query string is prepared by the Query Parser and the query string is searched with the index. The matching entries from the indexing table are ranked and are sent as the result to the user.

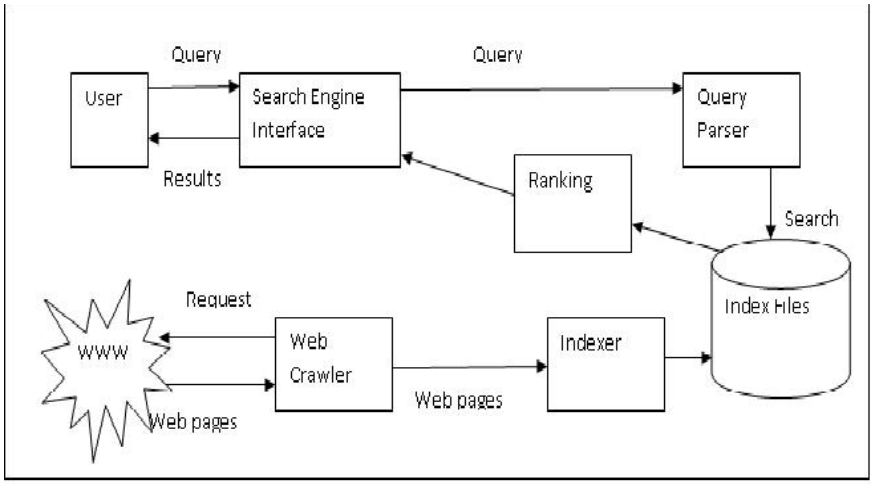

Fig. 1. Architecture of Web Search Engine

Web pages are filtered using Term Frequency-Inverse Document Frequency (TFIDF) scores for that page. TFIDF is calculated as the product of Term Frequency (TF) of a term t in document d and Inverse Document Frequency (IDF) of a term t.

$$TFIDF_{td} = TF_{td} * IDF_t \quad (3)$$

Term Frequency (TF_{td}) of document d and term t is given as the ratio of the number of occurrence of a term in a web page to the total number of words in that page.

TF_{td} = No. of term t in document d / Total no. of words in document d (4)

Inverse Document Frequency(IDF_t) of a term specifies the uniqueness of a document. It is given as a ratio of the total number of documents to the number of documents containing the term.

IDF_t = log [Total no. of documents (N) / No. of documents with term t] (5)

The higher value for IDF_t specifies that the document is unique as the term is present in few documents. The lower value specifies that the document is not unique.

3. META SEARCH ENGINE

Meta Search engine receives request from the user and sends the request to various search engines. The search engines check their indices and extract a list of web pages as links and pass the result to the Meta Search Engine. The Meta Search Engine receives the links, applies few algorithms, ranks the results and finally displays the result. The Meta Search Engine architecture is shown in Fig.2.

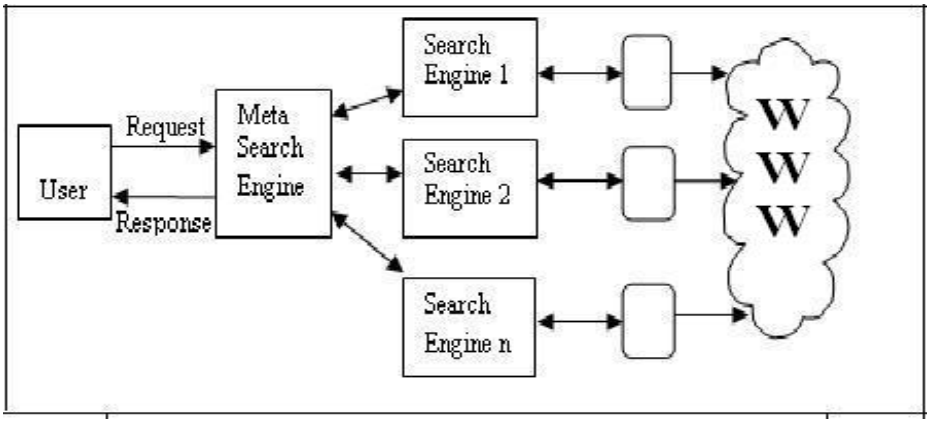

Fig. 2. Meta Search Engine Architecture

4. METHODOLOGY

Web Search Engines [10] such as Google, Yahoo, MSN and Vertical Search Engines[11] such as Yahoo Health, webMD, MSNHealth and Meta Search Engines [12] such as DogPile, MetaCrawler, MonsterCrawler are analyzed. A Multiple Search Engine [13] is also analyzed in which a Search Engine can be selected from a list of search engines and queries can be sent to it.

Four topics Cancer, Diabetes, Paracetamol and Migraine are searched in the 9 search engines. The results are classified under categories Relevant, Reliable, Redundant and Removed based on the retrieved web page contents. 20 results from each of the 9 Search Engines are considered for each topic and are given in Table 1. The interpretation of the results is shown in Fig.3.

Table 1. Relevance, Reliability, Redundant and Removed details for Search Engines

Search Engine	Relevant	Reliable	Redundant	Removed
DogPile	18.75	16.25	1.75	0.5
Google	19.25	19.25	0.25	0.25
Meta Crawler	17.75	14.25	4	0
Monster Crawler	19	14	1	0
MSN	18.25	12.5	1.5	1.25

MSN Health	1	1	0	0
WebMD	17	17	0.25	0
Yahoo	19.25	17.25	2.75	0
Yahoo Health	11	11	0	0

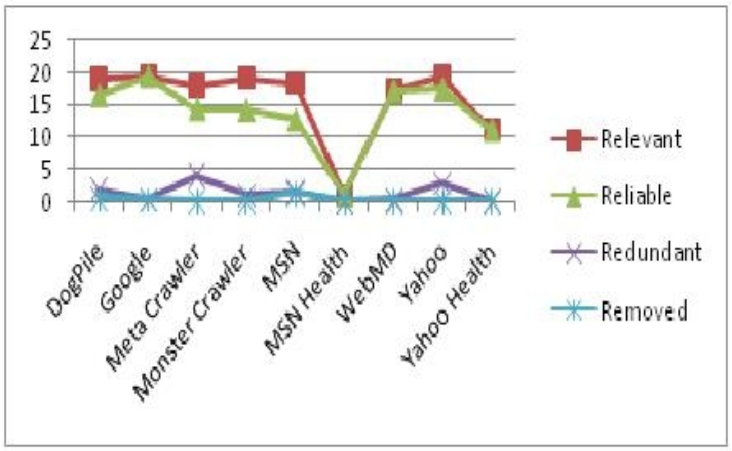

Fig. 3. Search Engine's Relevant, Reliable, Redundant and Removed details

The precision values are calculated using the ratio of relevant documents to the retrieved documents and is presented in Table 2. The precision details are also plotted in a graph and are shown in Fig. 4.

Table 2. Precision value table

Search Engine	**Precision**	**Search Engine**	**Precision**
DogPile	0.94	MSN Health	0.25
Google	0.96	WebMD	0.85
Meta Crawler	0.89	Yahoo	0.96
Monster Crawler	0.95	Yahoo Health	0.55
MSN	0.91		

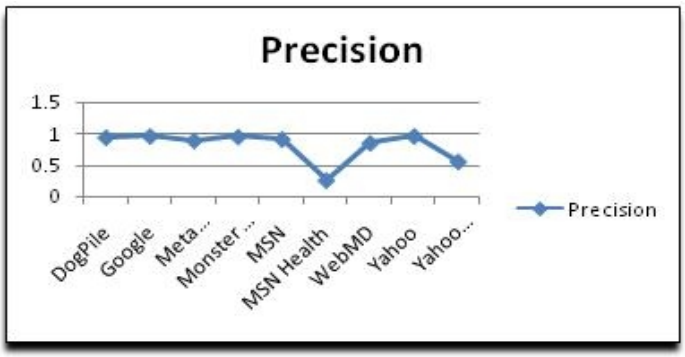

Fig. 4. Precision for various Search Engines for the given topics

The user searches for financial information, healthcare information and so on in the web. The user is expected to remember various search engine names to extract necessary information. The search engines also have a biased view of presenting the documents and the user may miss the necessary information. Hence the facility to get efficient domain related information on multiple domains is a need of the hour. Multi-Domain Meta Search Engine [14] sends queries to various search engines and consolidates the results from the search engines.

5. PROPOSED MULTI-DOMAIN META SEARCH ENGINE

We propose a user-friendly Multi-Domain Meta Search Engine that sends search queries to various search engines and to retrieve results from them. Various cases of selection of search engines are proposed in this paper.

The basic architecture of the Multi-Domain Meta Search Engine is given in Fig. 5. In the first model the Meta Search Engine was designed to send queries to Search Engines such as Google, Yahoo, AltaVista and AskJeeves. The queries were formed for a specific domain by adding the domain name as part of the search query. An interface for the Meta Search Engine was formed consisting of push buttons to select the domain and text box to enter the query string. The query string to be sent to the search engines are formed with the + symbol applied on the string entered and the selected domain name.

The second model of Multi Domain Meta Search Engine aims at providing an efficient information retrieval on a particular domain for the user by accessing various Vertical Search Engines.

The result of the query can be shown as a list of results from each search engine in individual windows, or as a list of results in different frames of a same window or as a single ordered list of results.

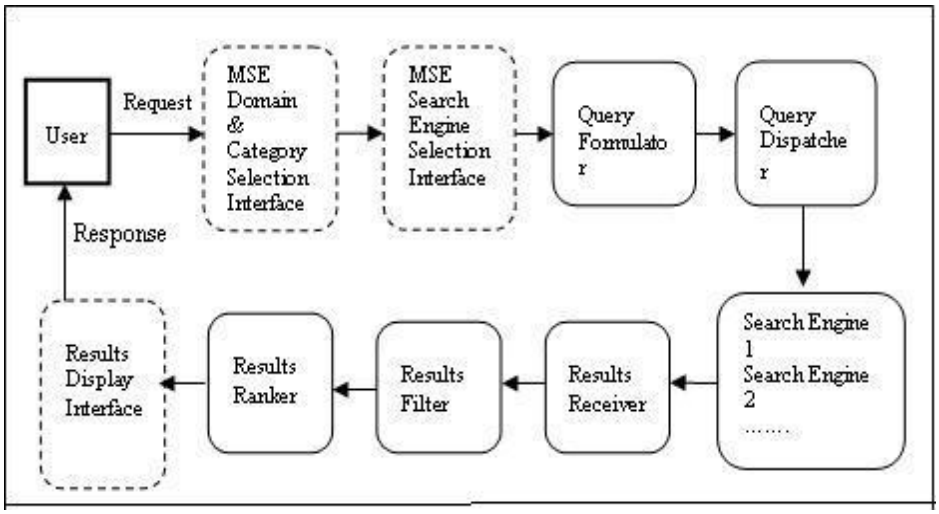

Fig. 5. Multi-Domain Meta Search Engine Architecture

6. MULTI-DOMAIN META SEARCH ENGINE OPERATIONS

The Multi-Domain Meta Search Engine consists of a user interface for domain selection, a user interface for search engine selection, query formulator, dispatcher, results receiver, filter, ranker and display of results.

6.1. User Interface – Domain Selection

The Domain selector presents a list of domains in this interface to the user. The user is allowed to

enter the request as a query. The user can directly move to search or can select the domain from which the information is requested.

6.2. User Interface – Search Engine Selection

The Search Engine Selection Interface presents to the user a list of Search Engines that are appropriate for the selected domain. Vertical Search Engines if available for a specific domain is also listed here. The user has the option to select one or more of the search engines. If the user has not selected the search engines then all the listed search engines are selected for further processing.

6.3. Query Formulator and Dispatcher

As different Search Engines follow different styles for the representation of the query search string, different search query strings are generated for a given user input. The query strings are then sent to various search engines to extract desired results from the search engines.

6.4. Results Receiver

The user generally searches only the first and second pages of a search result and also the most relevant results are presented in those pages. The Search Engines provide the best results also in the top few pages. Hence 25 results from each Search Engine are selected as the results from the Search Engines.

6.5. Filtering and Ranking Results

The results from a single search engine are found to be having redundant links and removed links during some of the searches. These links are removed from the results set from each of the search engine results in the first step. All the results are populated and the redundant links across the different search engine results are eliminated. The ranking used by the different search engines is used to rank the results.

6.6. Results Generation in different windows

The results from the various search engines are displayed in different windows. The user has the freedom to look into the results given in various windows. The advantage of this method is that the user can select as many Search Engines as needed. There is no need for the ordering and ranking of results. The disadvantage of the method is that the user has to switch between windows for effective usage of the system.

6.7. Results Generation in multiple frames of same window

The results from various search engines are displayed in various frames of the same window. The output window is designed to display results from 4 search engines simultaneously. This method is faster to use but the user has to make the decision about the result links to be visited. In this method also the ordering and ranking of the results is not done as the result from the various search engines are used as it is.

6.8. Results Generation in same window

Alternatively the ordering process is performed and the ranked list is displayed along with the rank and the name of the search engine that has produced the link.

7. RESULTS

The user is presented with a list of domains as shown in Fig. 6 and is allowed to select any one of the domains. The user is then presented with a set of Search Engines to be selected based on the selected domain. When the user selects the Medical Domain, a list of search engines Health Finder, Health Line, Web MD, Omni Medical Search and Pub MD is presented to the user. Any number if search engines can be selected from that list for the medical domain as shown in fig.7.

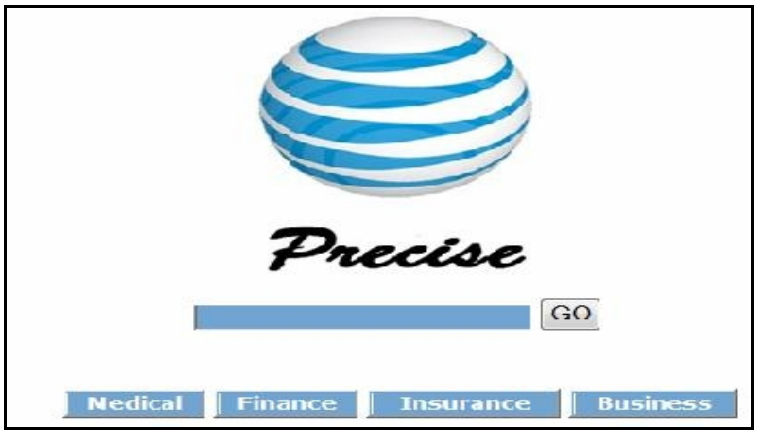

Fig. 6. Domain Selection Interface of Multi-Domain Meta Search Engine

Fig. 7. Search Engine Selection Interface of Multi-Domain Meta Search Engine

The result for the searches on Cancer, Diabetes, Paracetamol and Migrane is given in Fig.8 and the precision details are given in Fig. 9.

It can be seen from the figures 8 and 9 that the Multi-Domain Meta Search Engine's performance is better than the performance of individual search engines.

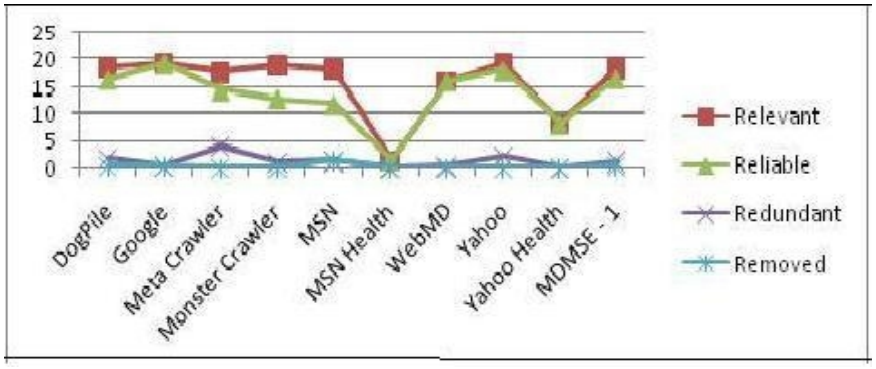

Fig. 8. Relevant, Reliable, Redundant and Removed data of Multi-Domain Meta Search Engine

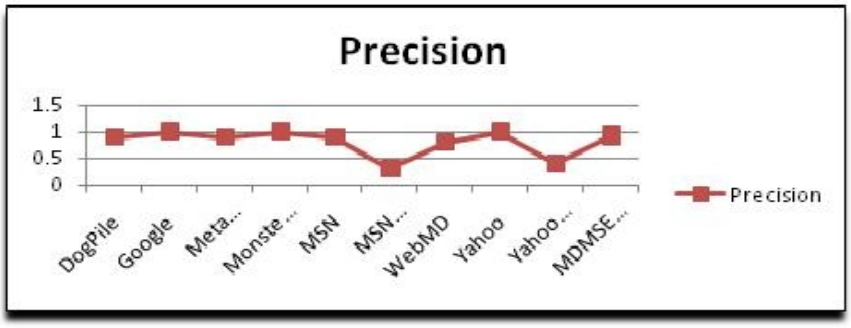

Fig. 9. Precision for MDMSE and various Search Engines for the given topics

8. CONCLUSION AND FUTURE WORK

The Web Search Engine, Vertical Search Engine and Meta Search Engine features are presented in this paper. Multi-Domain Meta Search Engines that provide an efficient information retrieval on various domains using various Search Engines are also presented. Few Search Engines are tested for the relevancy, reliability, redundancy and availability of search results for few topics. Domain selection and search engine selection user interfaces are also presented. A query interface is designed to send user's search query to the various search engines and the resultant links are consolidated to display the results to the user.

The system can be extended to process the content of the web pages in addition to processing of the links.

ACKNOWLEDGEMENTS

The authors wish to thank Ms.L.R.Rita, project student, Department of Computer Science, Madras Christian College, Chennai, India for her contribution towards development of the Multi-Domain Meta Search Engine.

REFERENCES

[1] Margaret H Dunham & Sridar S. (2007), *Data Mining*, Pearson Education.

[2] Raymond Kosla, Hendrik Blockeel. (2000), "Web Mining Research: A Survey", *SIGKDD Explorations*, Volume 2, Issue 1, pp 1-15

[3] Jeyaveeran N., Haja Abdul Khader A., Balasubramaniyan R. (2009), "E-Learning and Web Mining: An Evaluation", *Proceedings of the 2nd International Conference on Semantic e-Business and Enterprise Computing*.

[4] Rajashree Shettar, Rahul Bhuptani, "A Vertical Search Engine – Based on Domain Classifier", *International Journal of Computer Science and Security*, Volume (2): Issue (4), pp 18 - 27.

[5] Gang Luo, Chunqiang Tang, (2008), "On Iterative Intelligent Medical Search", SIGIR '08, *The 31st Annual International ACM SIGIR Conference Singapore*, July 20 - 24, 2008, pp 3 – 10.

[6] Gang Luo, (2008) "Intelligent Output Interface for Intelligent Medical Search Engine", *Association for the Advancement of Artificial Intelligence, 2008, Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp 1201 – 1206

[7] Ilic D., Bessell T.L., Silagy C.A. and Green S.(2003) "Specialized Medical search-engines are no better than general search-engines in sourcing consumer information about androgen deficiency", *Human Reproduction* Volume 18, No.3 pp 557 – 561.

[8] Gang Luo, (2009) "Design and Evaluation of the iMed Intelligent Medical Search Engine", *ICDE '09 Proceedings of the 2009 IEEE International Conference on Data Engineering*, pp 1379 – 1390.

[9] Kwok-Pun Chan,(2007) "Meta search engine". *Theses and dissertations. Paper 232*. http://digitalcommons.ryerson.ca/dissertations/232

[10] General Search Engine web sites: http://google.com, http://yahoo.com, http://search.msn.com

[11] Vertical Search Engine web sites: http://health.yahoo.net, http://www.webmd.com, http://health.msn.com

[12] Meta Search Engine web sites: http://dogpile.com/, http://metacrawler.com, http://monstercrawler.com

[13] Ryen W.White, Mathew Richardson, Mikhail Bilenko,(2008) "Enhancing Web Search by Promoting Multiple Search Engine Use", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 43 - 50*

[14] Minnie D., Srinivasan S.,(2011) "Meta Search Engines for Information Retrieval on Multiple Domains", *Proceedings of the International Joint Journal Conference on Engineering and Technology (IJJCET 2011)*, Gopalax Publications & TCET, pp 115 – 118

Authors

Ms. D. Minnie, M.C.A., (Ph.D in Computer Science – registered in 2009), Head In-Charge, Department of Computer Science, Madras Christian College, Chennai, India. Presented papers in National and International Conferences.Chairperson/Member in various Academic boards such as Board of Studies, Academic Council, Board of Examiners, Expert Committees. 2 decades of teaching experience and 6 years of research experience.

Dr. S. Srinivasan, M.Sc.(Maths), M.Tech (CSE), Ph.D.(CSE), Head, Department of Computer Science and Engineering. Director (Affiliations), Anna University of Technology Madurai, Madurai, India. No of Ph.D. students supervised/registered: 8. Presented papers in National and International Conferences.Chairperson/Member in various Academic boards such as Board of Studies, Academic Council, Board of Examiners, Expert Committees. 2 decades of teaching and research experience.
Membership in Professional bodies: ISTE, CSI