Meta Search Engines for Information Retrieval on Multiple Domains

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

[1]Madras Christian College, Chennai, India
Email: minniearul@yahoo.com
[2]Anna University of Technology Madurai, Madurai, India
Email: sriniss@yahoo.com

Abstract: **A Web Search Engine searches for information in the World Wide Web. The number of web resources increases every day but the user is often unable to get the exact information due to the different page ranking techniques followed by individual Search Engines. Meta Search Engines solve this problem to a certain level by using more than one search engines. A Vertical Search Engine is used to provide the user with results for queries on a particular domain such as Medical, Insurance and etc. This paper proposes three Multi Domain Meta Search Engines that facilitate efficient Information Retrieval on multiple domains. These Search Engines combines the functionality of both the Meta Search Engine and the Vertical Search Engine. The Search Results of the Meta Search Engines are found to be better compared to usage of an individual Search Engine.**

Index terms - **Information Retrieval, Web Search Engine, Vertical Search Engine, Meta Search Engine**.

I. INTRODUCTION

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from the World Wide Web. Web Mining consists of mining the Content, Structure and Usage of web [9].

A web search engine is a search engine designed to search for information on the WWW and returns a list documents in which the search query’s key words are found. Web Search Engines are classified as Crawler or Spider-based Search Engine, Directory-based Search Engine and Link-based Search Engine. The search engines are further classified into general purpose search engine such as Google that helps the user to search for anything and Vertical Search Engine.

Vertical Search Engine or Domain specific search engine or “Vortal” indexes the web pages that are specific to a particular domain to provide efficient search results on that domain. The filter component of Vertical Search Engine classifies the web pages downloaded by the crawler into appropriate domains [8]. Medical Search Engines such as “MedSearch” facilitate the user to get information on medical domain [2], [3]. Vertical Search Engines such as iMed creates search queries for the user by interacting with the user [1].

Different Search Engines follow different page ranking techniques and hence only 20% of the results are same for a specific search query. The user may fail to receive the necessary information. This gives rise to the use of Meta Search Engines [6]. Meta Search Engine accepts a search query from the user and sends the search query to a limited set of Search Engines. The results are retrieved from the various search engines and they are combined to produce a result set and given to the user. Users almost never search beyond 50 webpage results given by any Web Search Engine [4] and hence only they can be considered for further processing.

Information Retrieval is the science of searching for documents. Precision and Recall are two important metrics for retrieval mechanisms. Precision specifies whether the documents retrieved are relevant and Recall specifies whether all the relevant documents are retrieved.

$$\text{Precision} = \frac{\text{Relevant and Retrieved}}{\text{Retrieved}}$$

$$\text{Recall} = \frac{\text{Relevant and Retrieved}}{\text{Relevant}}$$

II. WEB SEARCH ENGINE ARCHITECTURE

A web search engine operates, in the following order: Web crawling, Indexing and Searching. Its architecture is given in Fig.1.

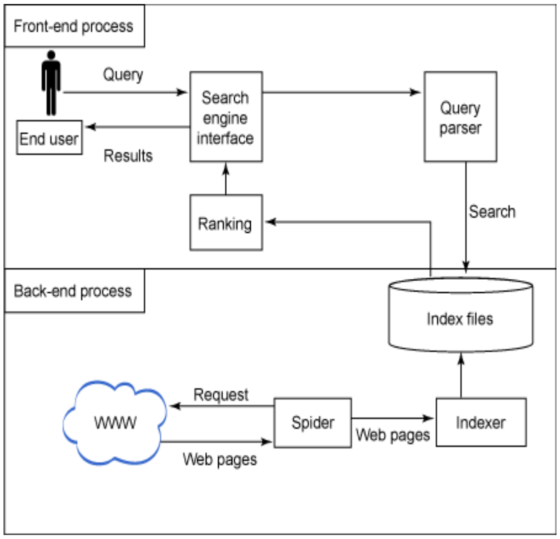

Fig.1. Architecture of Search Engine

Web search engines work by storing 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 directory-based search engine allows human editors to specify the documents that are to be included in the search space.

Web pages are filtered using TFIDF (Term Frequency-Inverse Document Frequency) 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.

$$TFIDF_{td} = TF_{td} * IDF_t$$

Term Frequency (TFtd) 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.

$$TF_{td} = \frac{\text{Count of term t in doc d}}{\text{Total no. of words in doc d}}$$

Inverse Document Frequency 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.

$$IDF_t = \log \frac{\text{Total no. of documents (N)}}{\text{No. of documents with term t}}$$

The higher value for IDFt specifies that the document is unique as the term is present in few documents. The lower value specifies that the term is present in many documents and hence the document is not unique.

Search engine indexing collects, parses, and stores data in an index data base to facilitate fast and accurate information retrieval. When a user enters a query into a search engine, the engine examines its index and provides a listing of best-matching web pages according to its criteria.

III. METHODOLOGY

The features of Web Search Engines [11] such as Google, Yahoo and MSN are analyzed. The Medical Search Engines [12] such as YahooHealth, webMD and MSNHealth are also analyzed to understand the features of vortals. Meta Search Engines [13] such as DogPile, MetaCrawler and MonsterCrawler are analyzed for understanding the properties of Meta Search Engines. A Multiple Search Engine is also analyzed in which a Search Engine can be selected from a list of search engines and queries can be sent to it [10].

Four topics Cancer, Diabetes, Paracetamol and Migraine are identified to be searched in the nine web search engines. The search results are classified under the categories Relevant, Reliable, Redundant and Removed based on the retrieved web page contents. 20 results from each of the nine Search Engines are considered for each topic and it is found that nearly 50% of the results are relevant. The reliability of the results is difficult to be accepted and also there is a presence of redundant results. Removed pages are also included in the result set.

The user is forced to remember many search engine names and addresses to get efficient information. It is also difficult to consolidate the information from those pages. The specific web sites are found to be having a biased view of presenting documents. Hence a Meta Search Engine for Multiple Domains is proposed here. The Vertical Search Engines and Meta Search Engines are analyzed to consolidate on the properties of Multi-Domain Meta Search Engine.

A Meta Search Engine interface is designed to send search queries to Web Search Engines, to Vertical Search Engines and also to Meta Search Engines. The queries are sent as it is for Vertical Search Engines and are modified and sent for general Search Engines.

IV. PROPOSED MULTI-DOMAIN META SEARCH ENGINES (MDMSE)

We propose Meta Search Engines that send search queries to various search engines and to retrieve results from them.

Various cases of Meta Search Engines are designed and analyzed in this study. The basic architecture of the first model of the Multi Domain Meta Search Engine is shown in Fig. 2.

In the first model the Meta Search Engine is designed to send queries to Search Engines such as Google, Yahoo, AltaVista and AskJeeves. The queries are formed for a specific domain by adding the domain name as part of the search query. An interface for the Meta Search Engine is formed with push buttons to select the domain and text box to enter the query string. The query string is formed by adding domain name with + symbol.

The second model of Meta Search Engine was designed to send queries to few Meta Search Engines.

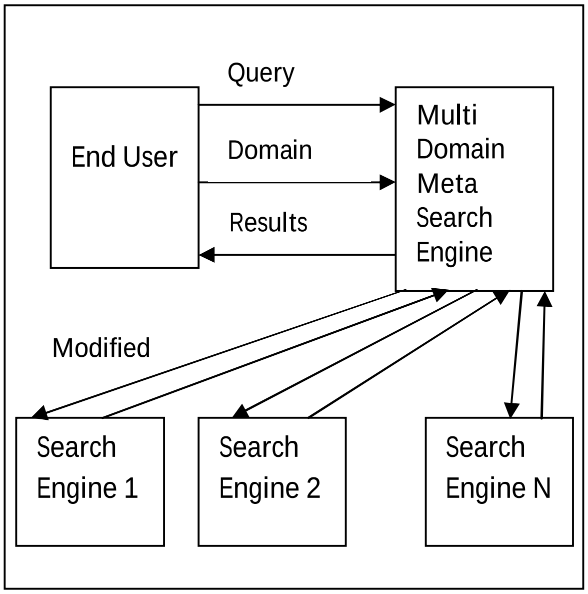

Fig.2. Basic architecture of Proposed MDMSE Model 1

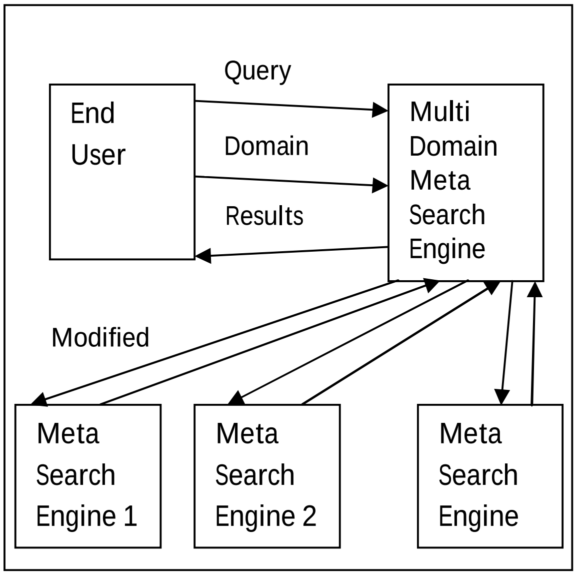

Fig.3. Basic architecture of Proposed MDMSE Model 2

The same interface used in the previous model is used for this case also. The search results from the Meta Search Engines are combined to produce the result. The basic architecture of the second model of Multi Domain Meta Search Engine is shown in Fig 3.

The third model of Multi Domain Meta Search Engine aims at providing an efficient information retrieval for the user by accessing various Vertical Search Engines and its basic architecture is shown in Fig.4.

The Multi Domain Meta Search Engine sends the search query to various Vertical Search Engines and retrieving results from such "Vortals". This creates an efficient and user friendly environment for the user to access the internet for efficient information retrieval on a domain such as medical, finance and insurance.

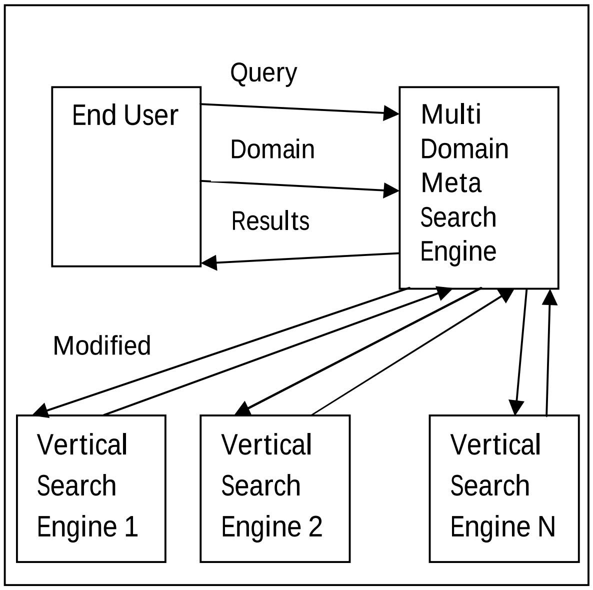

Fig.4. Basic architecture of MDMSE Model 3

V. PROPOSED MDMSE OPERATIONS

A. User Interface

The user is allowed to select a domain in which the information is required. The user is also presented with a set of Vertical Search Engines that are specific to that domain. All the listed Vertical Search Engines are selected if the user doesn't select them. The user is given the option to select the Vertical Search Engines that are to be used by the Meta Search Engine for that search.

B. Query Generation

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

C. Sending Queries to Vertical Search Engines

The generated queries are sent to the selected Vertical Search Engines with the help of programs written in JavaScript.

D. Receiving Results from Vertical Search Engines

The user generally searches only the first and second pages of a search result. 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.

E. Results Generation in multiple windows

The results from the various search engines are displayed in different windows.

F Results Generation in same window

Alternatively the received results are compared and the following techniques are applied to generate the set of web pages that are displayed as results to the user.

- Remove redundant results
- Remove outdated results
- Rank the web pages

VI. FUTURE WORK

Search Engines assume that users are capable of creating appropriate search queries and it is prove as not true in many cases. The users can be helped in forming an intelligent search query. An interactive module can be designed to accept the user's query and domain of search and then to ask the user few questions to narrow down on the formation of an efficient and improved search query that has to be approved by the user to be sent to the Search Engines.

The user can also be allowed to include new Vertical Search Engines to which also the query is to be sent by the Multi Domain Meta Search Engine.

VII. CONCLUSION

This paper analyses three Multi Domain Meta Search Engines that provide an information retrieval of information on various domains for the user using various Search Engines that are already available. The Web Search Engine, Vertical Search Engine and Meta Search Engine features are also presented. Few Search Engines such as Yahoo, MSN, Google, YahooHealth, MSNHealth, WebMD, DogPile, MetaCrawler and MonsterCrawler are tested for the relevancy, reliability, redundancy and availability of search results for few topics. A query interface is designed to send user's search query to the various search engines and to display results to the user.

ACKNOWLEDGMENT

The authors wish to thank Miss. L.R. Rita, for her tireless efforts in implementing and testing the Multi Domain Meta Search Engine Models as a part of her final semester project work.

REFERENCES

[1] Gang Luo, "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, 2009.

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

[3] Gang Luo, "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, 2008

[4] Ilic D., Bessell T.L., Silagy C.A. and Green S., "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, 2003.

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

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

[7] Margaret H Dunham & Sridar S, *Data Mining*, Pearson Education, 2007

[8] 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.

[9] Raymond Kosla, Hendrik Blockeel, "Web Mining Research: A Survey", *SIGKDD Explorations*, July, 2000, Volume 2, Issue 1, pages 1-15

[10] Ryen W.White, Mathew Richardson, Mikhail Bilenko, "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 (SIGIR 2008),* July 20-24, Singapore

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

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

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