Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

(Special Issue for ITECE 2016)

A User Centered Perspective Framework for Web Search

Nikita K. Shukla[1], Jemi M. Pavagadhi[2], Manilal B. Kalariya[3]

[1]*Computer Engineering, AVPTI, Rajkot,*
[2]*Computer Engineering, AVPTI, Rajkot*
[3]*Computer Engineering, AVPTI, Rajkot*

Abstract — **The Internet, and Specially the World Wide Web playing an important role of the current information age because of the huge amount of information and the large number of users who get access to this information. Data fusion on the web refers to the merging of the ranked document which is a unified single list; those are retrieved in response to the user query by more than one web search engine. Metasearch engine perform it and their merging algorithm utilize the information present in the ranked list of retrieved documents provided to them by the search engine, such as the ranked position of the retrieved documents and their retrievals scores. This paper investigates information retrieved by web search engine and provides effective and usable information to user.**

Keywords- Metasearch Engine, Data Fusion, Information Retrieval, Page Rank, Crawler, Dampster Shefer's Theory of evidence

I. INTRODUCTION

A Search engine is a tool that allows people to find information from World Wide Web. Search engines are constantly building and updating the index to the World Wide Web. They do this by using spider and crawl the web and fetch the web pages. Web search engine differ from one another in three ways – Crawling Search, speed, Frequency of updates and relevancy analysis.

Search engine has queried based information retrieval system (IR) that index and retrieve web documents. For a single search engine it's impossible to index the entire web for resources. Web user doesn't have sufficient knowledge as well as time to select appropriate search engine to search document. A Metasearch engine is a solution to prevail over this limitation. Metasearch gathered result from different search engine and help user to get more effective search result.
Metasearch engine has two problems first selecting the best combination of search engines to gather the most appropriate sets. This approach is know as database selection, which includes also database ranking. And second is selecting the appropriate method to combine the rank order of the retrieved sets. This process is known as result merging. To overcome this limitation data fusion approach was considered. Data fusion is the procedure to incorporate multiple data and knowledge representing the similar real-world object into a consistent, accurate, and useful representation.

II. RELATED WORK

A search engine is a web-based Program that enables users to locate information on the World Wide Web. It is a software program that search documents and files for keywords and returns the results of any files containing those keywords.

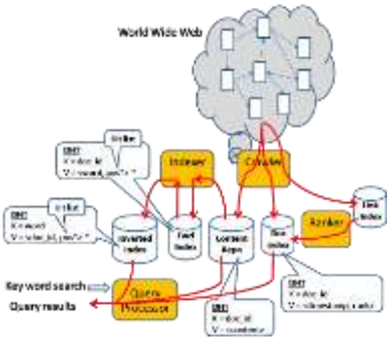


Figure 1. Working of Search Engine [10]

A basic search engine has a number of processing phases.

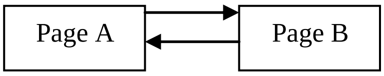
1. Crawling: to discover the web pages on the internet

2. Indexing: to build an index to ease query processing
3. Query Processing: Extract the most relevant page based on user's query terms
4. PageRank: to Order the result based on relevancy

Web search engines work by sending out a *Crawler* to fetch large number of documents as possible. Another program, called an *indexer*, then reads these documents and creates an index based on the terms contained in each and every document. Each search engine uses a proprietary algorithm to create its indices such that it returns significant results are returned for each *query*.

PageRank is one of the methods Search engine uses to determine a page’s relevance or importance. Number and Quality of link determines the importance of web page. Computation of pagerank can be done by three ways, first if the page has more back links considered to be more important, seconds is if page have more important point then page importance will increase.

Suppose Page A and Page B has same number of outgoing links,

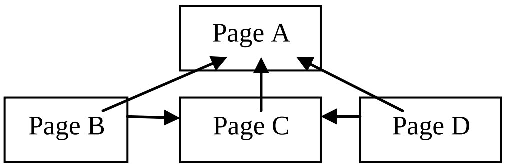


Then outgoing count is 1, i.e. C(A)=1 and C(B)=1, if page A has more important point then B and both have same outgoing link then Page A is more important than B. Third approach is if page have more outgoing links then that page sharing its importance to the child pages. Means importance of parent page propagated to child pages. Suppose we have four pages page A, B, C and D, and there is a link from B, C and D to A then the rank of Page A would be

R (A) = R (B) + R (C) + R (D)

Suppose Page B has a link to page A and Page C, Page C has a link to page A and Page D has a link to C and A. so the equation would be

R (A) = R (B) / 2 + R(C)/1 + R (D)/2



So in general page rank calculated by outgoing link equals to the page own rank score divide by its outgoing link L ( ). O_u denotes set of outgoing link for page u and B_u denotes set of pages that point to u. and page rank of u denoted by R (u) is defined by formula,

$$R(u) = \sum_{v \in B_u} \frac{R(v)}{|O_v|}$$

This equation indicates that more back link lead to larger page rank. Second R (v) is numerator indicate that page rank of u is increased as page v is more important (has a large R (v)). Third $|O_v|$ is denominator implies the importance of page is evenly divided and promulgate to each of its child pages. That means if $|O_v|$ increase then rank of the page will decrease.

III. DATA FUSION AND MERGING TECHNIQUES

Method 1 calculate the rank of the pages; Method 2 title and synopsis Method 3 Title, Synopsis and models data fusion using *dempster Shefer’s theory of evidence* Method 4 rank, title and synopsis of retrieved document Method 5 Reference Statistics of pages and at last Method 6 generate merged ranked lists by downloading and indexing the web documents.

Below figure explain functioning of Metasearch engine components. User requested query will be passed to MSEs web services which then redirected to multiple search engines. MSEs (Metasearch Engine) follow the raking algorithm which referenced different parameter to rank the pages. By using there parameters score of each page will be calculate, and to get more batter result data fusion will be applied which considered additional information to score the page. Result extractor extracts the result after that result merger merges the result and passed back to user.

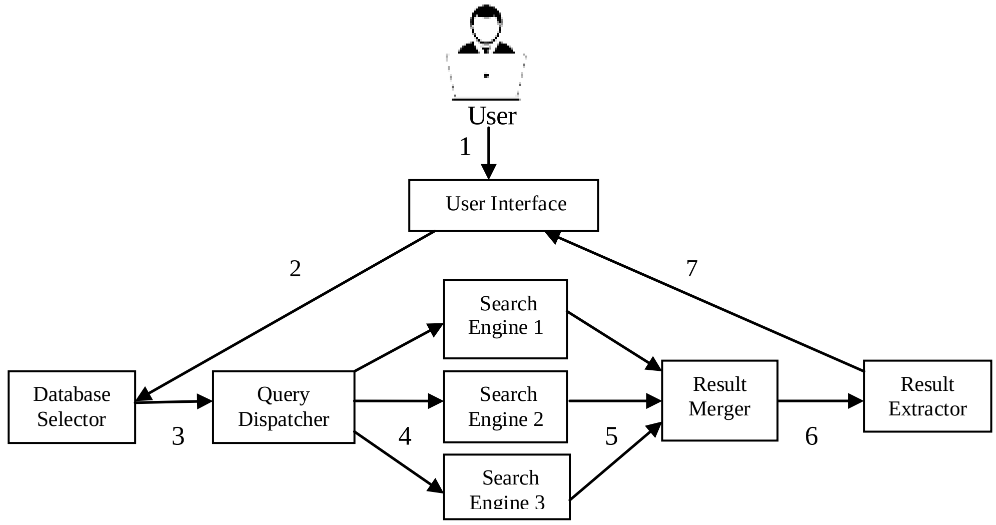


Figure 2. Architecture of metasearch engine components

3.1 Method 1: Merging using rank position

This method entirely includes information about the content of the documents, then search engine collect all the documents to determine the rank position, and the ordering is decided on the index terms of the documents. Hence this method is very simple as it required less information.

3.2 Method 2: Merging using the title and synopsis of the retrieved documents

This method calculates effectiveness of indexing title and summary of documents. Based on title and summary set of document will be indexed and represent these indexes in terms of vector. The weighting scheme used tf x idf [I] cannot be applied because the term idf (t) is equal to 0 when t=query. The documents are represented as d= {W1d, ……,wkd} here Wid is the weight of the ith index in term t in document d calculated using its term frequency tf (t, d). The similarity function used is $\sum_{t \in Q} w\,(t, d)$.

3.3 Method 3: Merging using Dempster shefer's Theory

A ceremonial approach can be introduce by defining merging algorithm which consider title and synopsis of collected documents and it's based on dempster shefer's theory of evidence. This will outcome to model of the data fusion process that is viewed as being equivalent to the aggregation of belief in uncertain reasoning.

3.3.1 Modeling Data Fusion on the World Wide Web using Dempster shefer's Theory of evidence

Documents produced by each involved search engine represented as a body of evidence defined in a stamp of discernment T. Each indexing term t E T corresponds to the basic proposal that 'the term t belongs to set of index term that index the documents contained in the list'. The bpa assigned to the merged list is calculated by combining the bpas of the individual lists, using the Dempster's grouping rule Bel(q) is used to rank the Web documents according to their estimated relevance to the query. Suppose that ranked list document l is combination of two individual ranked document l1 and l2 with respective bpas are m1 and m2. The bpa m associate to l is:

$$m\,(t) = m'\,(t) / K \text{ and } m\,(T) = m'\,(T) / K$$

Where,

$$m\,(t) = \begin{cases} m1\,(t)\,m2\,(t) + m1\,(t)\,m2\,(T) + m1\,(T)\,m2\,(t) & \text{if } t \text{ € } l1 \text{ and } t \text{ € } l2 \\ m1(t)\,m2\,(T) & \text{if } t \text{ € } l1 \text{ and } t \text{ € } l2 \\ m1\,(T)\,m2\,(t) & \text{if } t \text{ € } l1 \text{ and } t \text{ € } l2 \end{cases}$$

$m'(T) = m1(T)\, m2(T)$

where K is defined $K = m'(T) + \sum_{t \,€\, l1 \text{ or } t \,€\, l2} m'(t)$

3.4 Method 4: Merging Using Rank Positions, Title and Synopsis of the Retrieved Document

By introducing this method, aim is to investigating whether the combination of more information return by search engine (rank position, title and synopsis) leads to improvements in the effectiveness.

3.5 Method 5: Reference Statistics

Without use of integrated server it’s not possible to collect statistical database. This method having referential statistical database that contain relevant statistics for set of documents. This set either collection of searched document (about 12%) or complete new collection of documents.

3.6 Method 6: Merging By Downloading The Web Documents

This approach is based on the fact that since the entire document’s content is available the merging function can take this information into account and generate more effective merged ranked list.

IV. IMPLEMENTATION

Above six techniques perform data fusion by merging the list of documents which are locally stored, after being initially retrieved by search engine in response to a user query. These lists are passed so that each web document and its related ranked position, synopsis, title and URL extracted correctly.

V. RESULT

Experimental result which carried out in order to evaluate effectiveness of the merging methods, presented in this section. First of all, it was experiential that number of replica documents within the list of retrieved documents of all participating search engine is relatively small (5.23%).This lead to conclusion that there is a small overlapping among the pages that different search engines index, as it has already been recommended by similar studies. This indicates also that data fusion process raise breadth of the retrieval process, as the number of document return by the single search engine is less than the number of document return in merged list.

Table 1. Precision table for methods for each query

	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
Average	56.18%	62.49%	61.56%	62.95%	57.15	62.08%
%Change	0.00%	+11.23%	+9.57%	+11.21%	+10.08%	+10.50%
%Change	-9.05%	+0.66%	-0.84%	+1.02%	+0.72%	0.00%

VI. CONCLUSION AND FUTURE WORK

This paper explores that fusion operation effectiveness whether improved by merging methods proposed here or not. The experiment outcome indicates that merging method rank position is less effective than the merging methods *title* and *Synopsis*. But the combination of all the three methods (*title, synopsis and rank)* is more effective. *Downloading and analyzing* method is more sophisticated but slower approach to provide information to participating search engine. *Referential statistics* method is more effective in isolated server merging method. Finally merging function effectiveness of formal modeling data fusion improved by Dampster Shafer’s theory. More research required on ceremonial model of dam fusion procedure on web using Dampster Shafer’s theory of evidence. Finally the effectiveness of merging strategies can be improved by small number of queries. This research plays an important role in future on research of merging technologies.

REFERENCES

[1] Merging multiple search results approach for meta-search engines, By Khaled Abd-El-Fatah Mohamed, University of Pittsburgh 2004

[2] Personalized web search for improving retrieval effectiveness 2004 IEEE published by IEEE computer society

[3] Competition between internet search engines 2004 IEEE

[4] Merging Techniques for performing Data Fusion on the web CIKM’OI, November 5-10, 2001, Atlanta, Georgia, USA, Copyright2001 ACM

[5] Visualize Query occurrence in search result Lists 2005 IEEE
[6] Searching the web general and scientific information access IEEE Communication Magazine January 1999
[7] Dreilinger, D. & Howe, A. Experiences with Selecting Search Engines Using MetasearchI. ACM TOIS, 15(3), July1977, pp. 195-222
[8] Shafer, G. A Mathematical theory of evidence, Princeton University Press, 1976.
[9] The Google Pagerank Algorithm, Ian Rogers, IPR Computing Limited
[10] Search Engine Basics by Ricky Ho, MVB, March 4-10
https://dzone.com/articles/search-engine-basics
[11] System fusion for improving information retrieval system published by IEEE 2001