International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 8 (2017), pp. 2005-2013

https://dx.doi.org/10.37622/IJCIR/13.8.2017.2005-2013

Meta Search Engine with Semantic Analysis and Query Processing

Naresh Kumar
Assistant Professor, MSIT, New Delhi, India.

Praveer Singh
B. Tech. Student, MSIT, New Delhi, India.

Abstract

The increasing number of Internet users and digital immigrants has led to increase in enormous types of queries asked by the users. For finding the best results of the queries, abundant search engines and meta-search engines are being used with different and efficient result providing features. The existing meta-search engines uses various search engines for fetching the results but, do not emphasize on the semantic analysis of the query for identifying the best search engine suitable for the query of user. In order to overcome this limitation, a meta-search engine is proposed. The proposed meta search engine can improve quality of results through the use of semantic analysis and query processing. The end results comparison with existing meta search engine proves the proposed approach better than the existing meta search engines.

Keywords: Search engine, meta-search engine, semantic analysis, query processing, ranking.

1. INTRODUCTION

Presently, the Internet is playing a crucial role in day to day to life by accumulating prodigious and numerous natures of data [1] [2]. For getting information, a user enters query on the Search Engine (SE) interface [3] and expects the best possible result. SE responds by showing a list of web pages contains the expected information. But when

more than one SEs are used to performing the same task the diversity of results and coverage of WWW increases [4] which can enhances the satisfaction level of users. Due to this diversity and large coverage of WWW, Meta Search Engine (MSE) are getting more popularity.

A MSE is an information retrieval tool [5] used to retrieve the information from more than one search engine [6]. It is an easy way of getting huge information from number of SEs by using a single interface. But the existing MSEs are not able to understand the meaning of the user query. Moreover they are not able to respond the user query efficiently. This efficiency can be increased by introducing semantic analysis feature while implementing the MSE. So this paper proposed a new kind of MSE which can be used to overcome the above stated problem.

For developing the proposed MSE, the specialization of various search engines have been studied [3] and authors found Google, Bing and Yahoo as the best among available SEs. These SEs are used while testing the proposed approach.

The rest of paper is organized as: section 2 discusses in brief the related literatures and section 3 list the problem faced by MSEs. Section 4 explains proposed architecture and section 5 describe the experimental setup used in experimenting the proposed approach. Section 6 discusses the experimental achieved where as section 7 conclude the paper.

2. RELATED WORK

A new result merging algorithm for meta-search engine is introduced in [5] and the effectiveness of the merging algorithm is evaluated. The merging algorithm comprises of two algorithms – position merge algorithm and the titles and snippets merge algorithm. The position merge algorithm makes full use of original position information from each single search engine. For the titles and snippets merge algorithm, there is the need of downloading documents, index them and according to a kind of similarity function the similarity between the query and documents is computed. It concludes that the better merging algorithm could improve the quality of searching.

Study in [7] shows a Meta Search Engine to organize the results obtained from different search engines using ranking and clustering. The meta-search engine takes care of the relevancy and presentation of the search results and provides better results than the existing meta-search engines. It uses the relevancy calculator that calculates a relevancy score of the web pages returned by the search engines. A cluster generator module is implemented to generate clusters of same range of relevancy score web pages. The results show that the proposed meta-search engine gives high relevant results, removes duplicate links and performs clustering.

A heuristic approach called genetic algorithm for result merging in meta-search engine is proposed in [8]. The purpose is to identify the most relevant document according to the user from the huge amount of documents. It is based on the average weight of the document in selected search engine that is used in a fitness function. This concludes that the highest fitness valued documents are put in the top position in the merge list and considered as the most relevant results to the users' query.

In [9] ranking of retrieved URLs for meta-search engine is discussed. The work involves designing method for ranking retrieved results from different search engines for effective search results to increase reliability of meta-search results. Here each result is given a position in each search engine and counted the presence of result in different search engines. At the end the rank is calculated with the help of position and the count value and the results are ranked accordingly.

3. PROBLEM FORMULATION

The main problems identified in [5, 7, 8, 9, 10,11] are listed below:

i. MSE that performs clustering of documents and calculates the relevancy of documents deals with the performance deficiencies due to larger space required and time complexity.

ii. An implemented rank merging algorithm for MSE can provide more high quality results than general SE on average but MSE may not get better result than ordinary SE for any query as the meaning of users' query is not analyzed. Moreover they are not providing their own ranking.

iii. The Identification of most relevant documents is achieved with the help of result merging genetic algorithm but the expertise of respective search engines is not included.

iv. Most of the SEs does not undertake the concept of relevancy.

The proposed MSE is an attempt to overcome the mentioned problems.

4. PROPOSED ARCHITECTURE

The above stated problems are solved by developing MSE with genetic algorithm. The architecture of proposed architecture is shown in Figure 1. The major components of architecture are: Query Processor, Knowledge Base, Semantic Analyzer, Page Ranker, Page Retriever and Page Merger. The descriptions of each of these components are as discussed below:

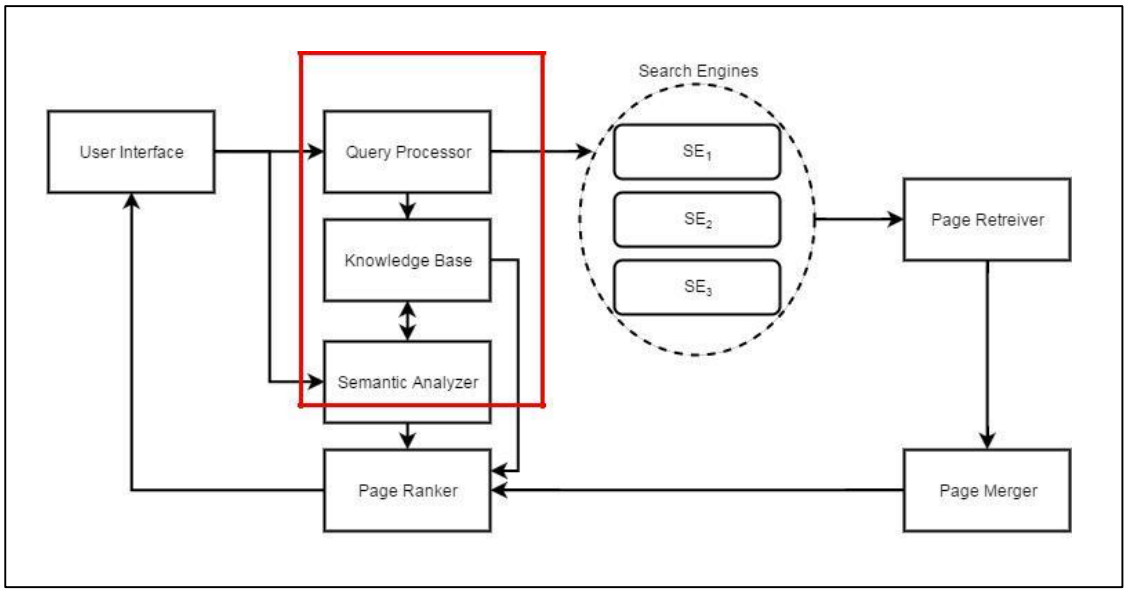

Figure 1: Proposed Architecture

a) **Query Processor**: It took query keywords entered by the user and passes them to participant SEs. Query processor works with the semantic analyzer to understand the meaning of users' query. It takes the user query as input with the help of user interface, converts the query into tokens and passes the tokens to the next module i.e. knowledge base.

b) **Knowledge Base**: The knowledge base is used to analyze the domain of the query entered by the user. This module is used in association with the dataset from which the keywords are matched. It acts as the warehouse of possible keywords that might help in identifying the domain of the query. Accordingly, the SE and its corresponding results are given additional weight.

c) **Semantic analyzer**: It took processed keywords which are processed by query processor and analyze them according to data set and gives output as the domain of the query. The domain of the query is achieved by analyzing the meaning and purpose of the query. The SE which gives best result for query of that domain will be chosen for the high priority results.

d) **Page Retriever, Merger and Ranker:** Page retriever fetches the documents from SE and passes the documents to page merger and ranker. For merging and ranking, a modified version of the genetic algorithm is used which is explained in Figure 2.

As compared to the genetic algorithm proposed in [8], this algorithm takes into the consideration the expertise of different SE by providing additional weight (t_m) to the SE, matching the domain of user's query. By doing this, the results of the additionally weighted SE get higher priority, hence, getting a better rank after compiling the search results. Hence, the relevant search results will be ranked better, thereby, improving the efficiency of the meta-search engine.

Algorithm: Modified Genetic Algorithm for re-ranking
Input: Let a set $S=\{U_q,D_L\}$ where $U_q$is the user query and $D_L=\{D_1,D_2,\ldots D_n\}$ is the list of document return by the underlying search engine .
Output: Single rank list $\{L_d\}$ of document after merging the result.
Method:
Begin
Step 1: Generate the random population of documents in the search space.
Step 2: Calculate the fitness of each individual of population using

A. First calculate the score of each individual of population using

$$Doc\ Score_{i,j} = \sum_{k=1}^{m} |L_i| - P_k + 1$$

Where $DocScore_{i,j}$ is the score of j^{th}document on i^{th} SE and $|L_i|$is the document returned by the i^{th} SE P_k is the position(s) of j^{th}document in the list returned by i^{th} search engine

B. After finding the document score the weight of each SE is calculated as:

$$w_i = \left(\frac{i}{n}\right)^a - \left(\frac{i-1}{n}\right)^a + t_w$$

Where n is the total number search engine, 1>a>0 is a real random number and $t_w$is the temporary weight that is added to the search engine matching the domain of the query.

C. Now we calculate Order Weighted Average and average for same

$$OWA_{i,j} = \sum_{j=1}^{m} w_i \times Doc\ Score_{i,j}$$

$$Avg(owa_{i,j}) = \frac{\sum_{r=1}^{X(U)} owa_r}{X(U)}$$

D. Fitness of an individual document is obtained as follows[11]

$$fitness(i) = \begin{cases} \beta, CP > fitness(i) || \max fit \\ \left|\frac{Avg(OWA)_i - CP}{\max fit - CP}\right|, otherwise \end{cases}$$

Where 1> β> 0 is a real random number, *max fit* is the maximum value of *Avg (OWA)* and CP is the cut point as *0.5×maxFit.*

Figure 2: Modified Genetic Algorithm

5. EXPERIMENTAL SETUP

The algorithm was implemented using Java, JSP and Servlet technologies. The queries from three domains were selected to test the proposed approach. These domains were Technology, Education and Health. The domain of the query is represented as the specialization of the query. The input dataset includes the set of results fetched from different SEs and output is the set of most relevant and unique search results. Three SEs were chosen depending on the percentage usages of SEs. The SEs was tested with queries of all domains, 30 times a day.

6. DISCUSSION OF RESULTS

Three different SEs- Google, Bing and Yahoo are chosen by authors to perform the experiment. The total number of results retrieved from each search engine, for the user's query, is shown in Table 1. The first column represents the name of the SE and succeeding column shows the number of results fetched for a particular query form that domain or SE. The number of search results returns by SE is in thousands. However the maximum numbers of search results that can be retrieved for free are limited to 100 per page per SE due to which authors were not able to increase the number of results in the testing of proposed approach.

For the three queries: "Javascript", "colleges in India" and "Hospital in India" the experiment was conducted and the obtained results in numbers are shown in Table 1.

Table 1: Number of results fetched by different SEs

Search Engine	Queries		
	Javascript	**colleges in India**	**Hospitals in India**
Google	100	100	100
Yahoo	65	66	56
Bing	68	72	65
	μ = ~230 results per query		

The mean of results (μ) was near about 230.

A total number of unique results in average is shown in Table 2. The first column represents the query and the next column shows the number of unique results obtained from Google, Yahoo and Bing.

Table 2: Uniqueness of results

Query	No. of Unique results
Javascript	165
colleges in India	150
Hospital in India	128
μ= 147	

For three queries mean a number of unique results are obtained are 147.

7. CONCLUSION

In this paper, the authors put forward a new method for presenting the result retrieved from the MSE by semantically analyzing the users' query by identifying the domain of the query and providing priority to the SE which has expertise in their domain. Although the MSE gives satisfactory results but it may lack in relevancy when some anomalous keyword in the query is identified. The authors has also compared the proposed MSE with other MSEs based on the number of SE used, Relevancy of Results, Rankin Criterion, Re-ranking of Results. The comparison is shown in Table 3.

Table 3: Comparison of Proposed MSE with other MSEs

Characteristic / Parameters	**Meta-Search Engines**					
	MetaCrawler	**WebCrawler**	**Excite**	**Dogpile**	**Gnome**	**Proposed MSE**
Number of search engine used	3	2	3	3	10	3
Result Relevancy	Moderate	Moderate	High	High	Low	High
Ranking Criterion,	Eliminate delicacy and display result	Based on lexical similarity	Three point scale	Simply collect the result and display	Page Ranking	Modified Genetic Algorithm
Re-Ranking	No	No	No	No	No	Yes

REFERENCES

[1] Patel B. et. al., "Ranking Algorithm for Meta Search Engine", in International Journal of Advanced Engineering Research and Studies, E-ISSN2249–8974, Vol. II, Issue I, Oct.-Dec., 2012, pp. 39-40.

[2] Nath R. et. al. "A new Approach for Implementation of Meta Search Engine using Ranking and Clustering ", published in Satyam, MSIT journal of research, ISSN: 2319-7897 vol. 1, No. 2, Jan - June 2013, pp. 11-14.

[3] Kanakam P., et al., "An Analysis of Exploring Information from Search Engines in Semantic Manner" International Journal of Advanced Engineering Research In Computer Science and Software Engineering,2014, IJARCSSE,ISSN : 2277128X pp. 793-801.

[4] Zonghuan W. et. al., "Towards a highly scalable and effective meta search engine", In proc. of 10th international conference on World Wide Web, Hong Kong, 2001, pp. 386-395.

[5] Yuan F. and Dong W., "An Implemented Rank Merging Algorithm for Meta Search Engine", International Conference on Research Challenges in Computer Science, 2009 (ICRCCS"09), Shanghai, 2009, pp. 191-193. DOI:10.1109/ICRCCS.2009.56.

[6] Shanfeng Z. et. al.. "Using online relevance feedback to build effective personalized Metasearch engine", In proc of IEEE, 2nd international conference on Web information systems Engineering (WISE'01), Kyoto, Japan, 2002, Vol.1, pp. 262 - 268.

[7] Kumar N. and Nath R., "A Meta Search Engine Approach for Organizing Web Search Results using Ranking and Clustering", in International Journal of Computer (IJC), vol. 10, No 1, pp. 1-7, 2013.

[8] Kumar J. et al., "Result Merging in Meta-search Engine using Genetic Algorithm", International Conference on Computing, Communication and Automation (ICCChhhA2015), Noida, 2015, pp. 299-303. DOI:10.1109/CCAA.2015.7148393.

[9] Patel B. and Shah D., "Ranking Algorithm for Meta Search Engine", International Journal of Advanced Engineering Research and Studies (IJAERS), 2012,ISSN: 2249-8974, pp. 39-40.

[10] Lu Y. et. al., "Evaluation of Result Merging Strategies for Metasearch Engines", 6th international conference on web information engineering, 2005, pp. 53-66.

[11] Kobayashi M. et al., "Information retrieval on the web," in ACM Computing Surveys, vol. 32, no. 2, pp. 144-173, 2000.

[12] Minakov et. al.," Development of Multi-agent Internet Meta-Search Engine: IT in Business (ITIB)", in International conference in St. Petersburg, June 14-17, 2005.