A Learning Based Search Engine Selection Technique

Kawaljeet Kaur , Richa Bansal

Abstract: A good search engine selection algorithm should identify potentially useful databases accurately. Many approaches have been proposed to tackle the database selection problem. Such as Rough approach, static approach and learning approach. To fulfill the demand of users and help them to be more effective on selecting relevant and useful search engines, this paper presents a search engine selection algorithm that is based on training query and demands of user query.

Keywords: Algorithm, Information retrieval, Metasearch engine, Personalization, Search engine selection.

I. INTRODUCTION

To get any kind of information, the Internet has become the major platform. In recent years, the web has become a huge source of information, which is mostly unstructured in the form of text or images. However, it is a challengeable task for the user to search for useful information from the huge amount of information on the internet[11] .The person all over the world poses queries using their favorite search engine to find relevant information. However, every search engine uses their own method for ranking the retrieved results[3].Search Engines are widely used for information retrieval, but there are lots of WebPages over the internet and a single search engine cannot cover all the web pages. A meta-search engine is a search engine that utilizes multiple search engines. A Meta search provide the solution for this problem, MetaSearch Engines (MSEs) are tools that help the user to identify relevant information. To perform a Meta Search, user query is sent to multiple search engines; once the search results returned, they are received by the MSE, then merged into a single ranked list and the ranked list is presented to the user. Most of the data on the web is in the form of text or image. A good database selection algorithm should identify potentially useful databases accurately. Many approaches have been proposed to tackle the database selection problem. These approaches differ on the database representatives, they use to indicate the contents of each database, the measures they use to indicate the usefulness of each database with respect to a given query, and the techniques they employ to estimate the usefulness.

Rest of the paper is organized as- Section-II describe about Meta Search engine, Section-III tells about the prior research, Section-IV shows Proposed work, Section-V describe experimental result, Section-VI compare the results, Section-VII is about the conclusion and Section-VIII describe future scope.

II. META SEARCH ENGINE (MSE)

A meta search engine is a tool that helps to locate information available via the WWW. It provide a single interface that enables users to search many different search engines, indexes and databases. Thus Meta search engines are capable of searching several search engine databases at once. Metasearch engines reduce the user burden by dispatching queries to multiple search engines in parallel [5]. Metasearch engine would collect the result from each engine, after comparing, analyzing, consolidating and deleting the repeat information, finally returns to users with certain format [1]. For each search engine selected by the database selector, the component document selector determines what documents to retrieve from the database of the search engine [2]. The top most documents having higher global similarity in the ranked list are returned to the user through the interface. In this survey, we concentrate on the search of text data. Query format is a list of keywords, called ''terms'' which provides the semantic to the documents. Ranking of the relevance documents is based on the weight of the query. There are a number of reasons for the development of a metasearch engine and we discuss these reasons below [8].

Increase the search coverage of the Web: A recent indicated that the coverage of the Web by individual major general-purpose search engines has been decreasing steadily. By combining the coverages of multiple search engines through a metasearch engine, a much higher percentage of the Web can be searched.

Solve the scalability of searching the Web: the problems associated with employing a single general purpose search engine will either disappear or be significantly alleviated. The size of a typical special-purpose search engine is much smaller than that of a major general-purpose search engine. It is also much easier to build the necessary hardware and software infrastructure for a special-purpose search engine. As a result, the metasearch engine approach for searching the entire Web is likely to be significantly more scalable than the centralized general-purpose search engine approach.

ISSN: 2278 – 1323
All Rights Reserved © 2014 IJARCET

Facilitate the invocation of multiple search engines: If a metasearch engine on top of these local search engines is built, then the user only needs to submit one query to invoke all local search engines via the metasearch engine. A good metasearch engine can rank the documents returned from different search engines properly.

Improve the retrieval effectiveness: Suppose that there is a special-purpose search engine for this subject area and there is also a general-purpose search engine that contains all the documents indexed by the special-purpose search engine in addition to many documents unrelated to this subject area. It is usually true that if the user submits the same query to both of the two search engines[8], the user is likely to obtain better results from the special-purpose search engine than the general-purpose search engine. This method has been shown to improve the retrieval effectiveness of the system As a result, if for any given query submitted to the metasearch engine, the search can be restricted to only special purpose search engines related to the query, then it is likely that better retrieval effectiveness can be achieved using the metasearch engine than using a general-purpose search engine.

III. PRIOR RESEARCH

This section surveys related work of data base selections. The objective of search engine selection is to improve efficiency as it would result in sending a query to only potentially useful underlying search engines. In paper [1,][3], utilizes the retrieved results of training queries for selecting the appropriate search engines for a specific user query.

The selection of the search engines is based on the value of relevance between user query and the training query.

Modeling Relevant Document Distribution (MRDD) [3] [10] is a static learning based approach, which uses a set of training queries for learning. With the help of training queries, it identifies all the relevant documents returned from every

search engine and arrives at a distribution vector for each relevant document. Similarly, it finds the distribution vector for each training query and calculates the average distribution vector is used to identify the appropriate search engines.

ReDDE [1] resource selection algorithm was proposed to estimates the distribution of relevant documents among available information sources for resource selection. ReDDE utilizes database size estimation and a centralized sample database (CSDB) that consists of the documents obtained by query based sampling. The CSDB is a representative subset of the centralized complete database (CCDB) which is the union of all the documents in available information sources. Since the CCDB is not available in the federated search environment, ReDDE uses the CSDB to simulate the property of CCDB.

Savvy Search engine [13], ProFusion[3] [14], is a hybrid learning approach, which combines both static and dynamic learning approaches. In the ProFusion approach, when a user query is received by the Metasearch engine, the query is first mapped to one or more categories underlying search engines. The query is mapped to a category that have at least one term that belong to the user query.

IV. PROPOSED WORK

Since there tends to be many similar queries in a real world meta search engine, the valuable information of past queries can help us provide better database selection results. In this section, we propose an algorithm, to utilize the valuable information to guide the decision of database selection. To propose an algorithm idea is taken by reference [1][3], in which the retrieved documents for each training query from all selected search engines are used to calculate the relevance between search engines and respective past query using top k document. The top k document is a rank merge list of documents from all search engines. By using the retrieved results of training queries, select the appropriate search engines for a specific user query.

In this algorithm more appropriate search engine contains more relevant information for the user query. The value of relevance depends on relevance between search engines and training queries and similarity between all training queries with the user query. The search engines with higher value of relevance being selected by the Meta search engine**.**

The selection of search engines is based on the value of relevance between user query and the search engine. Ranking in the search engines is carried out according to the value of relevance between search engines and user query. A higher the value of relevancy means that the search engine contains most relevant documents with respect to user query Uq . Therefore the search engine having higher value of relevancy is to be selected.

ALGORITHM:

Input: Let set A [TQ, Uq, s], where TQ is the number of training query, Uq is the user query, and S is a set of search engines.

Output: Sorted Order of topmost search engines.

STEP1: For each i^{th} training query, Rank the documents retrieved from all search engines into single ranked list

STEP2: Compute relevance [3] between search engines and training query.

$$\mathrm{Re}\,l\left({}^{s_j}\!\big/_{TQi}\right) = \sum_{\substack{\text{topD doc in the merged}\\ \text{rankedlist}}} \frac{\mathrm{Rel}\left(s_j/\mathrm{doc}\right)}{\mathrm{D}}$$

where D is a pre-defined number of top documents. if (doc € S) then $\mathrm{Re}\,l\left({}^{s_j}\!\big/_{\mathrm{doc}}\right) = 1$

Otherwise $Rel\left({}^{s_j}\!/_{doc} \right) = 0$

STEP3: Generate the ranked list R_q of documents returned from the search engines for the user query Uq

STEP 4: For each i^{th} training query TQ_i and user query Uq, compute the similarity using [13]

$$Sim(TQ, Uq) = \frac{\overrightarrow{TQ}\bullet\overrightarrow{Uq}}{|TQ|\times|Uq|} = \frac{\sum_{i=1}^{t} w_{ij}\times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}\times\sqrt{\sum_{j=1}^{t} w^2_{iq}}}$$

Where $w_{ij} = tf_{ij} \times idf_i$ and

$tf_{ij} = \dfrac{f_{ij}}{\max\{f_{1,j}, f_{2,j}, \ldots\ldots, f_{i,j}\}}$, Term Frequency (TF) is the weight of a term t_i in document dj is the number of times that t_i appears in document d_j, denoted by f_{ij}.

$idf_i = \log\dfrac{N}{df_i}$, Inverse Term Frequency (IDF) , N is the total number of documents in the system and df_i be the number of documents in which term t_i appears at least once.

$$w_{iq} = \left(0.5 + \frac{0.5 f_{iq}}{\max\{f_{1q}, f_{2q}, \ldots\ldots\ldots, f_{tq}\}}\right) \times \log\frac{N}{df_i},$$

w_{iq} is term weight of each term t_i in q

STEP5: Normalized the value of $Sim(TQ_i|Uq)$ using

$$MAXSIM_q = \max_i Sim(TQ_i|Uq)$$

$$CUTSIM_q = \alpha * MAXSIM_q$$

Where α is a constant

Normalized $Sim(TQ_i|Uq) =$

$$\left\{ \begin{array}{ll} 0 & if\ Sim(TQ_i|Uq) < CUTSIM_q \\ \dfrac{Sim(TQ_i|Uq) - CUTSIM_q}{MAXSIM_q - CUTSIM_q}, & otherwise \end{array} \right\}$$

STEP 6: For user query Uq For (each j^{th} search engine) compute

$$Rel(s_j / Uq) = \sum Rel(s_i / TQ_i) \times Sim(TQ_i / Uq)$$

STEP7: Ranked the search engine according to the value of $Rel\,(s_j|Uq)$. A larger value of $Rel\,(s_j|Uq)$ contain most relevant documents with respect to the user query Uq.

V. EXPERIMENTAL RESULT

The proposed algorithm is simulated in Matlab 2010b. Suppose we apply training query set TQ_i = [6 2 1 3 5 4; 6 5 1 2 3 4; 2 1 5 6 4 3] and user query Uq = [1 5 1 5 3 3 5 5] to the algorithm. The Meta search engine selects the document for each underlying search engines and for each training query, by random generation [3]. Finally a single merge list of all the document of each training query will be generated as per step-1. In similar way meta search engine generate a merge list for user query. And then Step-2 is applied to calculate relevance between each search engine and each training query.

To find the similarity between training query and user query Step-4 is applied. Result of Sim(TQ, Uq) is normalized by considering a suitable value of variable ranges from 0.1 to 0.8. In the last step relevance between search engine and user query is calculated, that is shown in the table1.

Similarity b/w TQ_i and Uq (i=1,2,3)	Normalized similarity on α=0.7	Similarity b/w S_i and Uq (i=1 to 5)	Sequence of selected search engine
0.6923	0.3939	1.5094	5
0.8462	1.0000	1.6760	2
0.8132	0.8701	1.6760	3
XXXXXXXXXXXXXXX		1.5750	4
		1.6977	1

Table 1.

VI. COMPARISON BASED ON RESULT

Following graphs shows the comparison between the existing algorithm [3] and proposed algorithm.

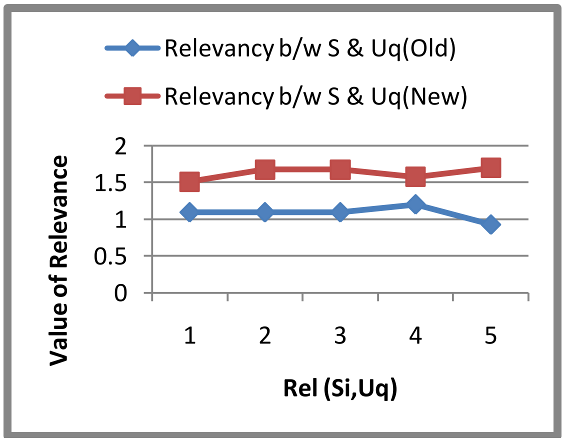

Figure 1 Similarity between Training query and user query.

ISSN: 2278 – 1323
All Rights Reserved © 2014 IJARCET

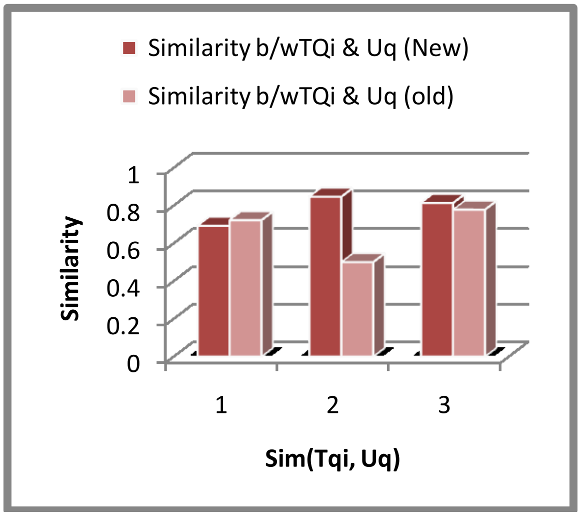

Figure 2 Relevance between search engines and user query.

VII. CONCLUSION

The value of relevance depends on relevance between search engines and training queries and similarity between all training queries with the user query. By computing all these values we find the order of selected search engines, having higher similarity with respect to user query.This paper implement an algorithm , that aimed to find more appropriate search engines those contains more relevant information with respect to the user query.

VIII. FUTURE WORK

Searching is a personal activity. Users have different interests, expectations and styles. To further improve Web searching, we must focus on the user, most immediately on how to identify and serve their goals. For example, the relative importance of waiting time, thoroughness, accuracy, and resource consumption all should be incorporated into determining where and how much to search. The resources of the Web are vast, but hardly limitless. Therefore to select the relevant search engine among the various search engines with respect to the user queries we optimize various algorithms like genetic algorithm, particle swarm optimization (PSO), magnetic optimization algorithms, simulated annealing and ant colony optimization etc.

.

REFERENCES

[1] SULEYMAN CETINTAS, LUO SI, HAO YUAN *"Learning from Past Queries for Resource Selection"* ACM CIKM'09, November 2–6, 2009

[2] DANIEL DREILINGER, ADELE E. HOWE *" Experiences with Selecting Search Engines Using Metasearch"* ACM Transactions on Information Systems, Vol. 15, No. 3, Pages 195–222, July 1997.

[3] R.KUMAR, A.K GIRI*"Learning Based Approach for Search Engine Selection in Metasearch"* IJEMR Volume-3, Issue-5, ISSN No.: 2250-0758, Pages 82-88, October 2013.

[4] G.TOWELL, E.M. VOORHEES, N.K. GUPTA, B.J LAIRD *"Learning Collection Fusion Strategies for Information Retrieval"* Appears in Proceedings of the Twelfth Annual Machine Learning Conference, Lake Tahoe, July 1995.

[5] ADELE E. HOWE AND DANIEL DREILINGER "*A Metasearch Engine That Learns Which Search Engines to Query"* American Association for Artificial Intelligence, AI Magazine Volume 18 Number 2, 1997.

[6] MANOJ M AND ELIZABETH JACOB *"Information retrieval on Internet using meta-search engine: a review"* Journal of Scientific & Industrial Research, Vol 67, pp. 739-746October, 2008.

.[7] HOSSEIN JADIDOLESLAMY *"INTRODUCTION TO METASEARCH ENGINES AND RESULT MERGING STRATEGIES: A SURVEY"* International Journal of Advances in Engineering & Technology, ISSN: 2231-1963, Nov 2011.

[8] YUAN FU-YONG, WANG JIN-DONG *"An Implemented Rank Merging Algorithm for Meta Search Engine"* International Conference on Research Challenges in Computer Science, IEEE, 2009.

[9] LUO SI AND JAMIE CALLAN, *"Relevant Document Distribution Estimation Method for Resource Selection" SIGIR '03*, July 28-Aug 1, 2003, Toronto, Canada. Copyright ACM, 2003.

[10] WEIYI MENG , CLEMENT YU , KING-LUP LIU *"Building Efficient and Effective Metasearch Engines"*ACM Computing Surveys, Vol. 34, No. 1, pp. 48–89,March 2002.

[11] XUE YUN, SHEN XIAOPING "Research on an algorithm of Metasearch Engine Based on Personalized Demand of Users" International Forum on information Technology and Applications IEEE 2010

[12] H. JADIDOLESLAMY " Search Result Merging and Ranking Strategies in Meta-Search Engines: A Survey" IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814, Vol. 9, Issue 4, No 3, July 2012.

[13] DRIELINGER, D. AND HOWE, "A Experience with Selecting search engine using Metasearch", (1997)

[14] Fan, Y. and Gauch, S. Adaptive agent for information gathering from multiple, distributed information source. In Proceeding of the AAAI Symposium on Intelligent Agent in Cyberspace, Page no.40-46, (1999)

Kawaljeet Kaur
M.Tech Student, RGEC Meerut

Richa Bansal (Guide)
Asst. Prof., CSE Deptt, RGEC Meerut

All Rights Reserved © 2014 IJARCET