Performance Analysis of Result Merging Techniques in Meta Search System

Sarita Yadav, Jaswinder Singh

Department of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology, Hisar, India

Abstract: Meta-search engines are searching tools that are mainly developed to enhance the retrieval performance of the World Wide Web finding tools. It increases the search coverage of the web [1]. These are based on result fusion technique which includes three major steps i.e., selecting the most comprehensive databases and ranking them properly, combining the retrieved results then merging the results in a sole list of documents using the most appropriate merging algorithm. At the combining stage i.e. result merging stage the Meta Search Engine uses different merging algorithms. These merging algorithms either merge results by finding similarity scores [4] between query and document or simply by their respective ranks in the result list returned by the various search engine. The motive of this paper is to study some of the algorithms based on whether they use rank or score for merging the results in final list. A new algorithm, named as hybrid algorithm is introduced that combines the efficiency of two algorithms. Depending on their relevancy their performance is analyzed.

Keywords: Meta Search Engine, Result Merging Techniques, World Wide Web.

Introduction

Information present on the World Wide Web is expanding constantly, making it impossible for a single search engine to index the entire web for a query. A Meta search engine is a solution to overcome this limitation. By merging multiple results from different search engines, a Meta search engine is able to enhance the user's experience for retrieving information, as effort required in order to access more materials is less. A Meta search engine is efficient, as it is capable of retrieving a large amount of data; however, ranks of websites stored on different search engines are different: this can draw an irrelevant documents. Other problems such as spamming also tremendously reduce the accuracy of the search. The fusion of search results from various search engines aims to handle this issue and improve the engineering of a Meta search engine. Data fusion is a problem solving technique [5] using the idea of integrating multiple data and knowledge representing the same real-world object into an accurate and useful representation. The expectation is that merged data is more informative than the original inputs. It is a key component in a Meta search engine. Once the results from various search engines are collected, the Meta search system merges them into a single ordered list. The effectuality of a Meta search engine is closely related to the result merging algorithm it employs.

I. Merging Algorithms

The result merger in Meta Search Engines uses separate methods for merging results. But all these methods are broadly classified on the basis of whether they use rank or score. Rank algorithms take into account the ranking of documents that are returned by the different search engines and further combine them into a single ranked list. Whereas, the algorithms using score, find the similarity score between the query and the content present in the document. Depending on the above classification, the three algorithms are chosen in this work for their performance analysis. These are as follows:

i. Concept similarity algorithm:

It [2] aims to find the concepts, where a concept is a keyword which has some relation with the documents and has some particular characteristic. For eg.,{data mining}, {data warehouse} , {data fusion} , etc are concepts.
It initially finds the most frequent word, by finding frequency of each word in a query.
Freq (w1) =Xw/Nk
Where,
w1, is element in a query
Xw, is number of W present in domain
Nk, is total number of element in domain. After finding the most frequent word, we have to find whether it belongs to the concept in a concept map. This can be done by finding the dependency of most frequent word with other word in domain.
Dep (w1:w2) = P(w1|w2) / P(w2)
Here, w1 and w2 are some frequent words.

Using this function, the dependency is found and thus concepts are formed. The concept is now considered as term and this term now belong to the document after the search. The concept has top N keywords extracted from document. An N X m matrix is generated and its row wise addition is taken, followed by arranging it in descending order. The top results that are above the threshold are returned to the user.

ii. Cosine similarity Algorithm(tf-idf):

This algorithm [2] deals with term frequency (TF) and inverse document frequency (IDF). The term frequency and inverse document frequency is a numerical statistic which shows that how important a word is to a document in a collection or corpus. The tf-idf value increases proportionally to the number of times a word appears in the document.

First we find term frequency which is the count of number of times the word appears in a document. This count is normalized to avoid bias towards the longer documents. It measures the importance of the term t within a particular document D. Thus we have the term frequency,

TF (t, D) = Number of term t in document D

The inverse document frequency (idf) is a measure of whether the term is commonly present or rarely present in all documents. Idf of a document is calculated by dividing the total number of documents by the number of documents containing the particular term, and then taking the logarithm of the quotient obtained,

IDF (t, D) = log (|D| / | t ε d: d ε D |)

The formula is given as:

$$TF-IDF(t,d,D)=\frac{\sum_{t\epsilon d}(TF\times IDF)_1\times(TF\times IDF)_2}{\sum_{t\epsilon d_1}\sqrt{(TF\times IDF)_1^2}\times\sum_{t\epsilon d_2}\sqrt{(TF\times IDF)_2^2}}$$

This value of tf-idf is calculated for each term in the document and the resulted values are passed to process the N X N matrix. Then the row wise sum operation is applied to find its relevance. Those whose values are higher than threshold are selected as search result and returned as final result.

iii. Modified Bayesian Method(Item-Item):

In this method [3] the position ranking of a result R is calculated based on result rank position. The formula is:

$P_r (R) = \Sigma r_i (R) / n$

Where,

n, is the number of search engine participating in result extraction

R, is the result returned by the search engine

$r_i(R)$, is the rank of returned results

Calculated P (R) is saved in Position Rank table for each search results of the search engine for ordering final result.

Then in second step, the occurrence probability of relevant and irrelevant result is calculated. This is done by using:

$P_{rel} = P_r (rel \mid r1 ...rn)$

$P_{irr} = P_r (irr \mid r1 ...rn)$

And further, an optimal relevance, Orel and irrelevance percentage, Oirr will be calculated. The formula is:

$O_{rel} = (P_{rel}*100) / n$

$O_{irr} = (P_{irr}*100) / n$

iv. Hybrid Algorithm

A new algorithm is introduced in this work. This algorithm combines the efficiency of concept similarity algorithm and cosine similarity algorithm. At last stage when both algorithms produce a final list of documents to be retrieved, the hybrid algorithm combines the lists to form a new list that is much efficient.

II. Merging Algorithms

There has been a great deal of work done in making Meta search engines a reality. Most of the results focus on the enhancement of the efficiency of the Meta Search results. Choon Hoong Ding, Rajkumar Buyya [6] proposed a guided Meta Search Engine, called Guided Google that serves as an advanced interface to the actual Google.com. The main goal of this application is to help ease and guide the searching efforts of novice web users towards their desired objectives. Weiyi Meng, Clement Yu, King-Lup-Liu [1] presented an overview of existing Meta search techniques concentrated on the problems of database selection, document selection and result merging. Nick Craswell, David Hawking and Paul Thistlewaite [7] introduced two techniques for merging search results: Feature Distance ranking algorithms and Reference Statistics. These techniques are found to be more effective than the existing ones. Javed A. Aslam, Mark Montague [9] proposed three solutions to the problem of Meta search: an optimal democratic voting procedures, the Borda Count; investigate a Meta Search model based on Bayesian Inference and a model for obtaining upper bounds on the performance of Metasearch algorithms. Yiyao Lu, Weiyi Meng, Liangcai Shu, King-Lup Liu [8] proposed the effectiveness of various algorithms experimentally using 50 queries from the TREC Web track and 10 most general purpose search engines. Danushka Bollegala, Yutaka Matsuo, Mitsuru [2] proposed a robust semantic similarity measure that uses the information available on the web to measure similarity between words or entities. K. Srinivas, V.Valli Kumari, A. Govardhan [3] proposed an approach based on the local rank and the position rank of the retrieved results. Jaswinder Singh, Parvinder Singh, YogeshChaba [4] performs the performance modelling of information retrieval techniques using the different similarity functions i.e. Jaccard, Cosine, Dice and overlap. Mohommad Othman Nassar, Ghassan Kanaan [5] observed factors that affect the performance of Data Fusion algorithms. All factors that affect the performance of data fusion algorithms are discussed and recommendations related to when and how to deal with these factors.

III. Methodology

So far, the studies that have been done for merging multiple search results provide different algorithms and approaches to merge these results. But the different merging functions are not compared to detect best algorithm. The primary motivation of this paper is to compare different merging approaches. The approaches are based on whether they use similarity measure for ranking results from isolated search engines or they just simply use the positional ranking methods. The comparative analysis of three algorithms chosen and a new algorithm named hybrid algorithm is done. Various steps of methodology followed are shown in fig.1.

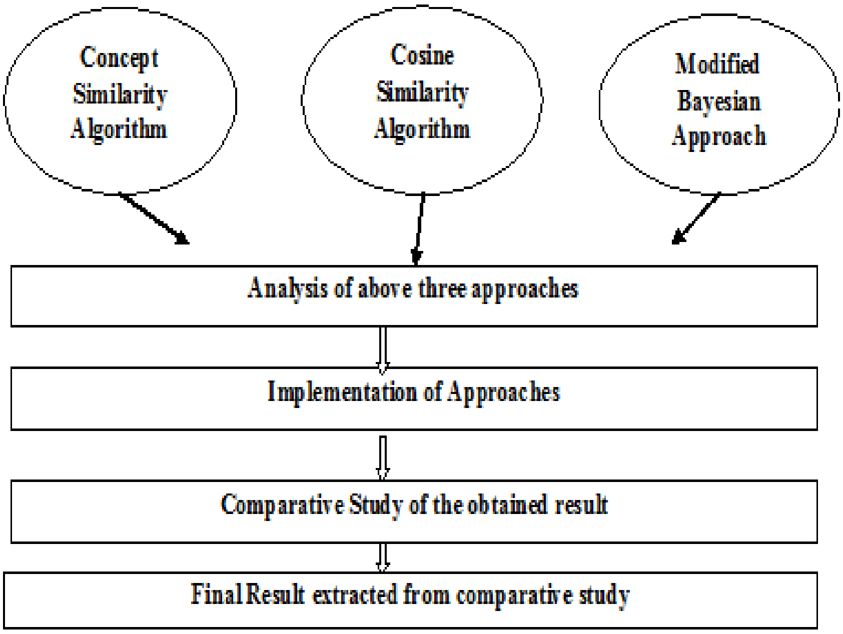

Concept Similarity Algorithm
Cosine Similarity Algorithm
Modified Bayesian Approach
Analysis of above three approaches
Implementation of Approaches
Comparative Study of the obtained result
Final Result extracted from comparative study

Fig.1 Methodology

IV. Results

The purpose of this work is to evaluate and compare different result merging algorithms. So, a database has been created and a GUI is developed. Each query is submitted using the GUI and according to the query the database is searched and results are returned performing different methods employed. The GUI is shown in fig.2.

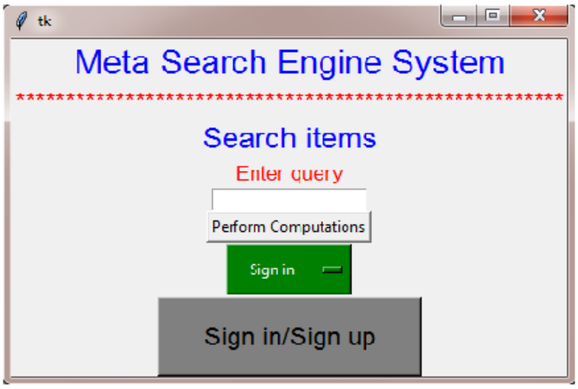
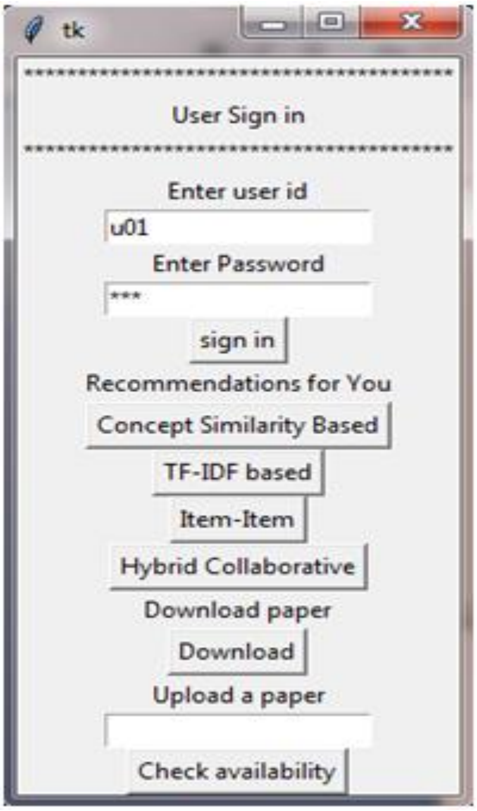


Fig.2 GUI-Search

Sign in is given to provide authentication. After sign in, a new window will appear as shown in fig.3. This window is for selecting the algorithm. When a query is entered and algorithm is selected then the top most results are fetched by applying the selected algorithm. If query 'recommendation' is entered and concept similarity method have been selected then the result will be as shown in fig4. Similarly, the results from all the methods are fetched for a single query. In the work, results have been fetched for various queries from all the methods**.**



Enter user id

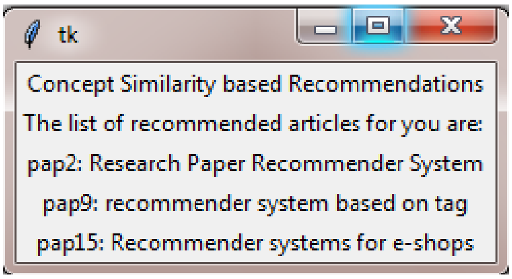
u01

Enter Password

***

Fig.3 GUI for selecting algorithm

When a query is entered in the text box of fig. 2 and algorithm is selected from fig.3 then the results are obtained. The GUI for obtained results is shown in fig.4.



Concept Similarity based Recommendations
The list of recommended articles for you are:
pap2: Research Paper Recommender System
pap9: recommender system based on tag
pap15: Recommender systems for e-shops

Fig.4 GUI-Result

V. Performance Analysis

In this section the performances of the approaches have been evaluated based on the different search results retrieved by the Meta Search Engine created in this work. Two search engines are also considered whose relevancy value is also given manually based on user's relevancy. The relevancy score to the topmost result stated as N, number of topmost results, is given to each result manually and according to the relevancy of individual results, the relevancy score is given to the output. Then the mean of this relevancy score is taken for analysis.

Table 1: Values for keyword = 'recommender'

Approach	N=10	N=20
Google	0.53	0.57
Yahoo	0.51	0.62
TF-IDF based	0.63	0.74
Concept Similarity based	0.67	0.69
Item-Item	0.43	0.53
Hybrid	0.71	0.73

The analysis from table 1 showed that the most ranked results are generated for the hybrid algorithm having relevancy score 0.71 for N=10 and 0.73 for N=20. But when score algorithms (concept similarity and tf-idf) are compared with rank (item-item) algorithm, the score algorithms outperform the rank algorithm with concept similarity having score of 0.67 for N=10 and 0.69 for N=20 and tf-idf have score 0.63 for N=10 and 0.74 for N=20. Whereas, rank algorithm (item-item) has score 0.43 for N=10 and 0.53 for N=20. Similarly other keywords are processed and analyzed.

VI. Comparative Analysis

In this section the comparison of score algorithm and rank algorithm along with the search engines is plotted. For query 'recommender' the graph is plotted for the relevancy value calculated from the output of different results by different algorithms. The graph is shown in fig5.

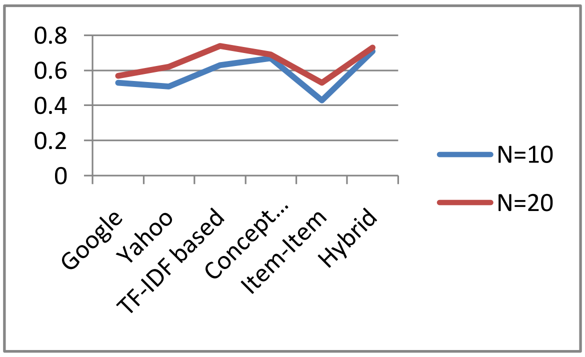


Fig.5 Comparison Analysis for keyword='recommender'

For some other keyword ‘item’, the graph is as shown in fig.6.

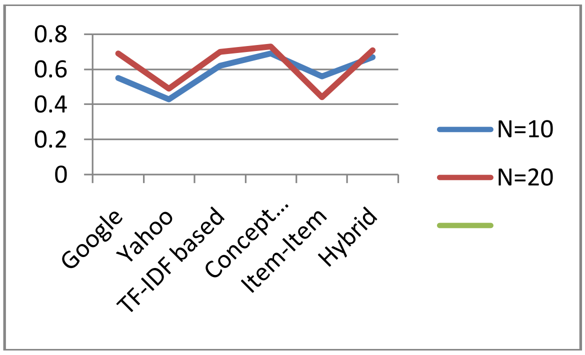


Fig.6 Comparison Analysis for keyword=’item’

Similarly other keywords are been used to compare the performance of these algorithms. The overall comparison is as shown in fig.7. The analysis shows that the score algorithms perform better than the rank algorithm. And among the score algorithm cosine similarity (tf-idf) algorithm performs much better

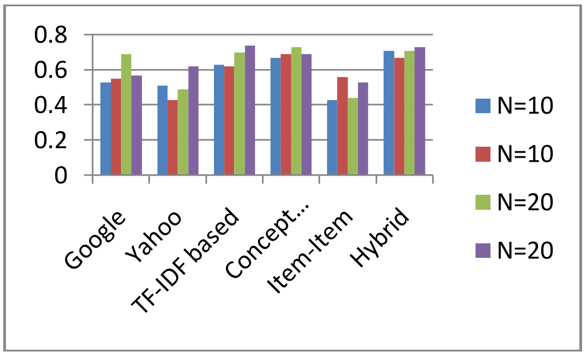


Fig.7 Comparative Analysis

Conclusion

This paper compares two methods on the basis of which Meta search engine fuses the result The Meta search Engines are capable to overcome the limitations faced by the normal search engines i.e. score or rank. Three algorithms using score or rank are compared and they are also compared with search engines and a newly designed method called hybrid method which combines the efficiency of both score algorithms. The result clearly indicates that the values for hybrid algorithms are higher. But comparing the score and ranking algorithms, tf-idf outperforms all other algorithm and even the search engines.

References

[1]. Weiyi Meng and King Lup Liu,” Building efficient and effective metasearch engines”, *ACM Computing Surveys*, Vol. 34,Issue 1,pp.1-50, 2002.

[2]. K. Srinivas, V. Valli Kumari and A. Govardhan,” Multi Similarity Measure based Result Merging Strategies in Meta Search Engine”, ACEEE *International Journal On Information Technology*, Vol. 3,Issue 2,pp. 90-97, 2013.

[3]. K. Srinivas, V. Valli Kumari and A. Govardhan,” Result merging using modified Bayesian method for meta search engine”, *In* the *Proc. of Conference on Information and Communication Technologies*, pp. 892-896, 2012.

[4]. Jaswinder Singh, Parvinder Singh, YogeshChaba,”Performance Modelling of information Retrieval Techniques Using Similarity Functions in Wide Area Networks”, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4, Issue 12, pp. 786-793, 2014.

[5]. Mohammad Othman, Ghassan Kanaan,” The Factors affecting the performance of data fusion algorithms”, *in the Proc. of International Conference on Management and Engineering*, pp. 465-470, 2009.

.

[6]. Choon Hoong Ding, Raj Kumar Buyya,” Guided Google: A Meta Search Engine and its implementation using Google Distributed Web Services” *http://arxiv.org/ftp/cs/papers/0302/0302018.pdf*, 1998.

[7]. Nick Craswell, David Hawking and Paul Thistlewaite,” Merging results from isolated search engines”, *Proceedings of 10th Australian Database Conference,Springer*, 1999.

[8]. Yiyao Lu, WeiyiMeng, LiangcaiShu, Clement Yu and King- Lup Lip, “Evaluation of result merging strategies for Metasearch engines”, in the *Proc. of 6th International Conference on Web Information System Engineering*,Vol. 3806, pp.53-66,Springer, 2005.

[9]. Javed A. Aslam and Mark Montague,” Models for metasearch”, *in the Proceedings of the 24th Annual International Conference on Research and development*, pp. 276-284,ACM, 2001.

[10].Hossein Keyhanipour, BehzadMoshiri, Maryam Piroozmand and Caro Lucas,” Web fusion : fundamentals and principals of a novel meta search engine”, *International Joint Conference on Neural Network*,pp.4126-4131,IEEE, 2006.

[11]. Biraj Patel and Dipti Shah,” Ranking Algorithm for Meta Search Engine”,*International Journal of Advanced Engineering Research and Studies*,Vol. 2,Issue 1,pp.39-40, 2012.

[12].Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka,” Measuring semantic similarity between words using web search engines”, *International world wide web conference committee*, pp. 757-766, ACM, 2007.

[13].Yuan Fu-Yong and Wang Jin-Dong,” An implemented rank merging algorithm for meta search engine”, *In the Proc.* of *International Conference on Research Challenges in Computer Science*, pp. 191-193.IEEE. 2009.