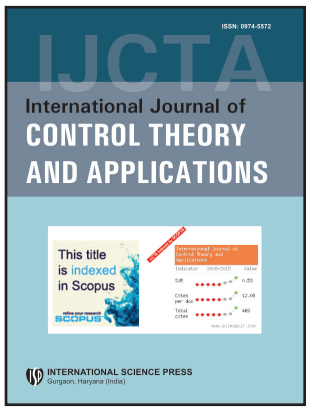

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 9 • Number 43 • 2016

Ranking Techniques Challenges and Solutions in Meta Search Engine

Naresh Kumar[a], Shourya Mittal[b] and Sakshi Kataria[c]

[a-c]Department of Computer Science & Engineering, MSIT, Janakpuri, New Delhi, India. Email: [a]narsumsaini@gmail.com; [b]mittalshourya97@gmail.com; [c]sakshi781996@gmail.com

Abstract: For the retrieval of useful information and to improve the search coverage, Meta Search Engine are being used widely. The efficiency of every Meta Search Engine depends upon the ranking and merging strategy employed to rank the Web Pages. Most of the Meta Search Engine provides results without its own ranking which leads to irrelevant or low quality of results. So, this paper discusses and presents various Meta Search Engines available in literature as well as available online. Further, the analysed Meta Search Engines were tested and examined by using hypothetical data. At the end, superlative algorithm to rank the search results was also discussed for some suitable parameters.

Keywords: WWW, Search Engine, Meta Search Engine, Page Rank, Ranking, Evaluation.

1. INTRODUCTION

With the increasing popularity and growth of Internet, gave a large boost to WWW. Initially, Internet was a collection of indexed pages used to navigate between specific files. The web pages were crawled according to web indexing, but large size of web make it insufficient to crawl completely.

As the time passes the size of web also increases from thousands of Web Page (WP) to billions of WP. At present, web has 4.77 billion WP approximately [17]. So, it becomes very difficult and challenging task to search the information from WWW. So, an automated software called search engine (SE) is used which helps the user in retrieval of information [4]. A study of [16] shows that any SE can cover only 16% of the entire web. Further, the expertization of different SEs are also different [9]. User lacks in both i.e. coverage and quality of information. So, a new platform called (Meta Search Engine) MSE came into existence.

Meta search engine is an effective way to cover a larger section of web by sending the query to various SEs [3]. MSE combines all the retrieved results to form a unified ranked list. It facilitates the user as large amount of information can be retrieved with same amount of exertion thus, saving time, providing quality results and better search performance. A MSE [14] does not contain its own database which is the major limitation of any MSE because most of the MSE ranks the results on FCFS basis. They does not provide any ranking on its own. Using specialized-purpose SEs [9], a MSE can solve the scalability of searching the Web. The effectiveness of a MSE depends on result merging algorithm [3].

In this paper, authors have discussed many ranking techniques used by MSE available in literature as well as available online. This paper has surveyed and analysed many research works done by different authors on different ranking techniques in MSE. Many ranking methods such as Advanced iral Ranking method [3], UMSE [5], Fuzzy Logic Implementation [6], Global KE algorithm [7] etc. have been discussed to focus on specific needs of users. Generally, MSE presents the results in two ways i.e. ranked list and clustered ranked list. But, authors of this paper have discussed only those MSE which provides the ranked results instead of clustered ranked results.

2. EXISTING WORK

The authors of this paper analysed the MSE stated in the information available in literature as well as available online. So, this section is divided into two parts i.e. literature survey and online available MSE.

A. Literature Survey

1. The Global KE Algorithm proposed in [7] personalizes the retrieved results by combining Geo KE Algorithm, Weighted KE Algorithm and URL Aware KE Algorithm. The SEs used - Google, Bing, Yahoo and Ask.

   The main features of proposed system are:

   (i) It considers geographical location of the user by assigning different values of G for ranking.

   (ii) User can limit similar domain pages by assigning the Domain Awareness constant (D) to retrieved results.

   (iii) It is improved process as user has the facility to give importance to SEs for retrieving results.

   (iv) Exploits regional information of webpage and user to give language familiar pages to the user.

   The proposed method is implemented in QuadSearch MSE and tested for the query "Athenian Acropolis" and compares GeoKE algorithm and KE algorithm by taking top ten results. Author has claimed that the introduced algorithm works efficiently for travel related queries and can be used to achieve better ranking according to the profile of user.

2. The combination of Position Merge algorithm (PMA) and Title/Snippet Merge Algorithm (TMA) was used to rank the retrieved results [1]. The results were presented to the user in decreasing order of final score calculated by combination of PMA and TMA. The authors checked the effectiveness of the proposed merging algorithm using the TREC-style average precision technique.

   As a result, authors claimed that merging algorithm can improve the quality of searching, but it is not necessary that MSE always get better result than the general SE for any query.

3. Fusion method in [8] used for ranking was based upon creation of user profile from the previous queries, SE, query identifiers, query terms and documents viewed and then ranking by Borda Adaptation model. It focuses on specific needs of user while querying. Also, a knowledge base is maintained from information of user, traces of links, etc.

   Fusion involves 2 phases – (i) Learning Phase for generation of user profile and (ii) Ranking Phase-gives score to a document in comparison to query. It also takes into account weight of individual SE for ranking and basic indicators such as recall and precision for testing the results of proposed method. The authors tested the method for information and technology related queries from dataset

"ClubWeb09". The experimental results proved that the proposed method is better than CombHMEAN and Borda models.

4. An approach proposed in [2] focuses on position of results on each SE for ranking of results in MSE. According to this, decreasing order ranks are assigned to the retrieved results out of the total results taken. The final ranks are computed by taking the ratio of sum of ranks of each result to the count of the result in each SE. Finally, the results are arranged in decreasing order of final ranks.

   For testing of proposed work, authors choose five search results and three SE and results proved to be quite promising.

5. Usearch MSE proposed in [4] considers two factors for ranking (i) similarity of search query with the description of retrieved results (ii) user satisfaction towards each SE. The pre-processing module generates two lists consisting of - (i) Common and (ii) Non-common results. The calculated scores of both lists and results are then ranked in decreasing order of scores by merging module. Final list is presented to user through the Interface module.

   The features of **Usearch** are:

   (i) It allows the user to invoke SE of his choice through interface module.

   (ii) Efficiency is increased by calculating the score of common and non-common results separately.

   The testing bed comprised of single, two and three word queries. At the end, authors concluded that top three result of each participating SE are always present in top twenty results of merged list.

6. UMSE proposed in [5] solves the problem of query translation and result merging in MSE. It uses a defined query syntax to solve the query translation problem to get more relevant information. User can invoke SE according to the choice and query to be searched through the Query Interface module (QIM). The query can be submitted to SE as well as user can search the query among the results available in "local repository". The results from SE are collected through "Result Collector". The information about every retrieved result is stored in local repository. Each SE is assigned a precedence by the user through (QIM). The merging algorithm compares whether the r^{th} ranked result of SE with priority p exists until $(r-1)^{th}$ ranked results of other SE. The duplicate results are removed from the merged list.

   Features of UMSE are:

   (i) User can submit long queries without having concern about the query syntax of SE.

   (ii) It provides off-line mode to allow user to search the query in local repository maintained over the previously downloaded results.

   (iii) In the offline mode user can find out the best SE, by changing their priority, and determining which SE returns the most relevant results.

7. The author in [6] focuses on ranking of results by fuzzy logic. The proposed algorithm, firstly finds count of participant SE in which the returned result is present and then maps the count to fuzzy value (i.e. if count = 3, then fuzzy value = 1; count = 2, fuzzy value = 0.5). Final results are presented to user in decreasing order of fuzzy value.

8. A Meta Search Engine named "iral" was proposed in [3] which make use of SEO parameters for ranking of retrieved results. Advanced iral ranking algorithm assigns equal weight to SEO parameters

such as title tag, meta description, meta keyword, snippet, meta expires, meta content, sitemap, image attribute, links, etc to calculate the value of rank parameter. Moreover, pages having synonyms to user query are also considered for ranking. Page rank is calculated by the product of weight and value of SEO parameter.

Features of this method are:

(i) Reduction in web spamming as a combination of SEO parameters is used for ranking.

(ii) Increased efficiency as SEs weights are assigned to each SE.

(iii) Modifies original PageRank algorithm for effective ranking.

(iv) Maintains Hash Table for the removal of redundant links by assigning unique hash values and for sending the sorted results to Page Ranker module for ranking.

(v) Knowledge Base sends queries to Dictionary.com to find out synonyms of user query thus eliminating the need of database.

(vi) It is proficient for single word queries.

The author has claimed accuracy of 0.48% (greater than that of Google (0.44%) and Bing (0.31%)) by considering the queries ”alcoholism” and “local computer shop” by 300 students.

B. Online MSEs

Nowadays, a large No. of MSE are being used. Some of them analysed by the authors of this paper are described as:

1. *Dogpile:* An intelligent MSE [11] organises results from Google, Yahoo, Yandex, Ask, Bing etc. and from audio and video content providers. The main features of Dogpile are links for White and Yellow pages, recent searches, IntelliFind, preferences, spelling corrections, statistics bars, search filter, favourite fetches, Toolbar etc. One of the unique feature of [11] is it provides search suggestions for any query. Moreover, it has an excellent user interface and produces results of relatively high relevancy than any SE.

2. *Excite:* MSE [10] has a customized user interface that serves a variety of services such asInternet portal- for displaying news and weather related updates , Web based email service, Instant messaging, Ticket selling- booking tickets for different events and venues, Education portal-that enables the user to search and apply for different degrees and online courses in America.

3. *Apocalx:* A MSE [12] based on results from Google, Yahoo and Bing. It provides the facility to retrieve results in four languages- English, Francis, Espanol and Italiano. The unique features of Apocalx are search options such as ApocalX recipes for food related searches, ApocalX Maps for searching the places. It also includes ApocalX reviews to get the reviews for different products, ApocalX countdown to set countdowns, etc. Moreover, it also allows a user to create his own countdown.

4. *IxQuick:* MSE [13] provides the stand-alone proxy service which allows the users to open search results via proxy. It incorporates “Star System” for ranking of results by searching results in eighteen languages using local SE. It manages user’s future search preferences by using a single cookie called “preferences”.

3. CHALLENGES

After studying the MSEs stated in literature survey and online MSEs, authors of this paper experienced some problems. They are summarized as follows:

1. Few mappings between User's Geographical location and language of the returned document in [7] may not be present in the database. Moreover, the authors have not specified the value of factor "G" when the user understands the language of the page returned as well as domain name and the user location match. The computational time required for ranking in GeoKE algorithm may be extremely large.
2. PMA used in [1], has lower accuracy. Further, the computational time required for calculating ranks is large making the user to wait a long for the final results. Furthermore, No. method is provided for calculating the weight factors such as W_{gp}, W_{gf}, N_{tf}, N_{pf}, N_{ap}, N_{af}, etc.
3. In [2], No. provision for ranking of pages having same score. It only takes into account the position of result on each SE. To increase the efficiency of ranking, authors haven't specified any method that compares query with the document description.
4. UMSE introduced in [5], needs user to have knowledge about the query format to submit the query to different SE.
5. Fuzzy logic implemented in [6] considers only the count value for ranking of results but documents having same count value will be ranked irrespective of their relevancy with the query. It neither takes into account the position of the result on SE nor the SEO parameters for ranking.
6. The coverage area of approach [3] is limited only to Google and Bing. Also, the proposed algorithm is not efficient for more than one word queries.
7. Borda Adaptation model used in [8], there is No. focus on semantic content of the query and on ontology, it only considers query terms as objects and retrieves results on the basis of matching of these objects with knowledge of user profile. Result retrieval time increases for query terms having similar meaning.
8. MSE specified in [11], [13] don't cluster web log data to discover similar access patterns, but produce the results by just combining retrieved results from different SEs. Thus, it lacks in grouping similar data and presenting to the user.

4. HYPOTHETICAL ANALYSIS

The study of literatures available in [1], [2], [6], [7] were tested manually for the query "Heart Transplant" by considering hypothetical values for various factors. For analysis purpose, five results were taken randomly from top 10 results of three popularly used SE *Google, Yahoo and Bing* as listed in Table II. The following hypothetical data was taken:

1. In [1], **for PMA,**

   Priority value: Google, W1 = 15; Yahoo, W2 = 10; Bing, W3 = 5

   For TMA,

   $$\alpha = 100, \beta = 20, W_{gf} = 10, W_{gp} = 5, K = 5, \theta = \gamma = 1$$

2. For [2], hypothetical values are given as: Considering top 10 links for all SEs and assigning position values as:

First link-10

Second link-9

Last link-1

3. For [7], D = 10, if result domain name is not repeated more than 2 times in result list

   D = 5, if result domain name is repeated more than 2 times in result list

   G = 1, if domain extension of result is same as user region and user understands language of page.

   G = 2, if domain extension of result and user's region match

   G = 3, if user understands the language of result page

4. G = 4, if domain extension doesn't say anything about location

   G = 5, if user doesn't understands the language of result page

   K = 10

   $e(i)$ = 10 for Google

   $e(i)$ = 9 for Bing

   $e(i)$ = 8 for Yahoo

5. In [6], Fuzzy values are assigned as;

   If count = 3, fuzzy value = 1; count = 2, fuzzy value = 0.66; count = 1, fuzzy value = 0.33; count = 0, fuzzy value = 0.

Table 1
Analysis of existing Meta Search Engine

S. No	*Features*	*Dogpile*	*Ixquick*	*Excite*	*Apocalx*
1	Location-based	No	No	No	No
2	Category Links	Yes	No	No	No
3	Front-end	Excellent	Good	Bad	Good
4	Shows Recent Searches	Yes	No	No	No
5	White page and Yellow Page Facilities	Yes	No	Yes	No
6	Result relevancy	High	High	High	Moderate
7	Clustering	No	No	No	No
8	Re-ranking	No	Yes	No	No
9	Research based	No	No	No	No
10	No of search engines used	3	Various(14)	3	3

Table 2
Original ranks of search results in SE

sE	*SR1*	*SR2*	*SR3*	*SR4*	*SR5*
Google	1	2	4	10	6
Bing	1	2	3	5	-
Yahoo	1	3	2	8	7

Table 3
Ranking of results in MSE

SR	*[1]*	*[2]*	*[6]*	*[7]*
SR1	1	1	1	5
SR2	3	2	2	4
SR3	2	3	3	3
SR4	5	5	4	2
SR5	4	4	5	1

6. The system proposed in [4] was analysed for different set of data. Top 10 links of three SE Google, Yahoo, Bing were considered and lists of common and non-common links were prepared. Rank was calculated for results using the following values:

   $Query_{occ} = 30$; $Keyword_{occ} = 20$; $Title_{occ} = 10$; $C1 = 50$; $C2 = 30$,

   (i) Google $C3 = 3$, $CC1 = 0.5$

   (ii) Yahoo $C3 = 2$, $CC2 = 0.4$

   (iii) Bing $C3 = 1$, $CC3 = 0.1$

   The results are available with the authors in tabular form.

   The non-common list for Google contains four results, for Yahoo, it contains five results and for Bing, it contains four results. The common list contains five results. Both the common and non common lists were ranked by MSE and returned to the user.

5. OBSERVATION

Authors took a query "Heart Transplant" and analysed it manually for most relevant results returned by Google, Bing, Yahoo. After that same query was evaluated on the literature available in [1], [2], [6], [7]. The results shows that MSE available in [1] provide most relevant result to end users as shown in Table III. Moreover, authors also analysed some existing MSE on some features listed in Table I. The rigorous analysis of the table shows that [13] provides more relevant results in real-time as user can define the query syntax for the SE to understand semantic content of query and to fetch more relevant results. Moreover, it provides various options for previous searches. It hides the IP address and cookies from SE, thereby provides privacy to user. The coverage is better than other stated online MSE as it ranks results retrieved from large number of SE.

6. SOLUTIONS

The feasible solution to above stated problems are listed as follows:

1. The ranking method of [2] only considers the position of result in SE. The efficiency of the ranking can be improved by using other techniques such as PMA, TMA or a combination of both etc.
2. The ranking method used in [6] can be improved by considering SEO parameters such as title tag, meta description, meta keyword, snippet, meta expires, meta content etc and thereby, checking the relevancy of retrieved documents with the query. Also, original page rank of the results can be taken into account for effective ranking.
3. Coverage area of [3] can be increased by incorporating more SE, to enhance the ranking of results. The efficiency of search results for more than one word queries could be improved by

employing semantic analysis techniques and ontology, thus understanding the context of the query.

4. An alternative approach to PMA algorithm in [1] is to use Query Dependent Ranking Algorithm applied in [15]. Also, the ranking can be done on the basis of calculating similarity between query and training query, by use of Artificial Neural Network.

5. Creation of named clusters in [11], [13] and grouping similar type of data would reduce the time required for user to get relevant information. For doing the same, K-Means or WEKA tool could be used for creation of clusters and appropriate algorithm can be used for ranking of these clusters produced.

7. CONCLUSION

It is very difficult to manage and present the information in an efficient way. Lots of literatures and online meta search engines are available for searching the information from the web. The authors have studied various MSE ranking algorithms as mentioned in section 2.1 and some of existing online MSEs in section 2.2. After complete analysis some of the existing challenges were listed. Also, some feasible solutions are provided. At last, [1] is considered to be the most superlative algorithm for ranking. Future work involves improvement of [2] by combining TMA with proposed algorithm. Incorporation of TMA will match the query with title/snippet of the result along with position of result in SE, resulting in effective results.

REFERENCES

[1] Yuan Fu-yong, Wang Jin-dong, "An Implemented Rank Merging Algorithm for Meta Search Engine", published in International Conference on Research Challenges in Computer Science, DOI: 10.1109/ICRCCS.2009.56, 2009.

[2] Mr. Biraj Patel, Dr. Dipti Shah, "Ranking algorithm for meta search engine", at Patel et. al., International Journal of Advanced Engineering Research and Studies, E-ISSN2249-8974.

[3] Jai Manral, Mohammed Alamgir Hossain, "An Innovative Approach for online Meta Search Engine Optimization", published in 6th Conference on Software, Knowledge, Information Management and Applications, Chengdu, China, Url: https://arxiv.org/abs/1509.08396, September 9-11 2012.

[4] Tarek Alloui, Imane Boussebough, Allaoua Chaoui, Ahmed Zakaria Nouar and Mohamed Chaouki Chettah, "Usearch: A Meta Search Engine based on a New Result Merging Strategy", Url: http://ieeexplore.ieee.org/document/7526966/, 2016.

[5] I. Anagnostopoulos, I. Psoroulas, V. Loumos and E. Kayafas, "Implementing a customised meta-search interface for user query personalisation", DOI: 10.1109/ITI.2002.1024655.

[6] Dr. Biraj Patel, Dr. Dipti Shah, Jay Patel, "Fuzzy logic implementation for ranking of search results in meta-search engine", at Patel et. al., International Journal of Advanced Engineering Research and Studies, E-ISSN2249–8974, Int. J. Adv. Engg. Res. Studies/III/IV/July-Sept., 2014.

[7] Leonidas Akritidis, Dimitrios Katsaros and Panayiotis Bozanis, "Effective Ranking Fusion Methods for Personalized Metasearch Engines", DOI: 10.1109/PCI.2008.31.

[8] Abdelbaki Issam, Benlahmar El Habib / LabrIn Elhoussin, "Result merging for Meta-search engine", DOI: 10.1109/SITA.2013.6560802.

[9] Amir Hosein Keyhanipour, Behzad Moshiri, Caro Lucas, "User Modeling for the Result Re-ranking in the Meta-Search Engines via the Reinforcement Learning", DOI: 10.1109/ISDA.2007.68.

[10] http://www.excite.com/,2016

[11] http://www.dogpile.com/,2016

[12] http://www.apocalx.net/,2016

[13] http://www.ixquick.com/,2016

[14] P. Vijaya, G. Raju, Santosh Kumar ray, "S-MSE:Asemantic Meta Searcg Engine Using Semantic Similarity and Reputation Measure", Journal of Theoretical and Applied Information Technology, 20th February 2014. Vol. 60 No. 2, ISSN: 1992-8645.

[15] Dilip Kumar Sharma, A.K.Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2670-2676.

[16] Naresh Kumar , Rajender Nath, "A Meta Search Engine Approach for Organizing Web Search Results using Ranking and Clustering", International Journal of Computer (IJC), ISSN 2307-4531.

[17] http://www.worldwidewebsize.com/