

# Document Clustering Approach for Meta Search Engine

**Dr. Naresh Kumar**

Assistant Professor, MSIT, New Delhi, India

**Abstract:** The size of WWW is growing exponentially with ever change in technology. This results in huge amount of information with long list of URLs. Manually it is not possible to visit each page individually. So, if the page ranking algorithms are used properly then user search space can be restricted up to some pages of searched results. But available literatures show that no single search system can provide qualitative results from all the domains. This paper provides solution to this problem by introducing a new meta search engine that determine the relevancy of query corresponding to web page and cluster the results accordingly. The proposed approach reduces the user efforts, improves the quality of results and performance of the meta search engine.

**Keywords:** Search engine, meta search engine, ranking and clustering.

## 1. INTRODUCTION

The gigantic size [1] of the web and the number of web pages is growing exponentially [2]. The information from such types of repository can only be retrieved by using some tools like search engines. As per a literature available in [3] every Search Engine (SE) has its limited search space and expertization [4] which may restrict the number of relevant web pages returned to user. Furthermore a study of [5] and [6] also indicates that coverage and the precision of different SE are diverse and restricted in nature. Therefore a single SE is helpless to fulfil all the requirement of the end user from all domains. This is the major reason due to which technology developer are using the concept of meta search engine (multiple SEs on single interface) [7], [8]. Meta Search Engine (MSE) is used to receive the information from many SEs concurrently [8], [9]. It receives the URLs from different SEs, delete the redundant results and present them to its users [8]. Efficiency of any retrieval system depends upon the relevancy and presentation of results to end user [5] where the existing MSE fails. Therefore this paper proposed an architecture to overcome the problem of relevancy and presentation of results. The proposed architecture will reduce the end user efforts for searching the results.

## 2. RELATED WORK

A MSE called Helios was proposed in [10]. Initially Helios was implemented by using eighteen SEs. This limit can be expanded as per demand. In it HTTP Retriever module had the responsibility of handling the network communications. It uses a dual PIV 2.60 GHz, 1.5



GB of RAM and 100 Mbps internet connection to implement the proposed approach. The performance of Helios was compared with wget. Nearly 600 results were retrieved in 12.4 seconds whereas Helios retrieved the same number of results in 4.6 seconds. According to the authors this approach can be used highly engineered open-source parallel meta-search engines and can be used in industrial environments.

A MSE was purposed in [11] where three SEs – Google, Yahoo and Baidu were used in

implementation. The position of the words and the snippets of the webpage were used to calculate the similarity of webpage. Top twenty results were selected to test the proposed approach. The results were tested physically by using the predefined criterion. TREC - style average precision was used to evaluate the results. At the end they claimed that most relevant results were appeared on the top of the returned resultant list.

Authors of [12], proposed a Multi domain MSE for effective presentation of results. It provides the facility of selection of specialized SEs. Relevancy, Reliability, Redundancy of results and accessibility of searched results were considered for performance measurement. The searched results were shown in the corresponding search engine window only. Finally the authors proved that the performance of proposed MSE is better than the individual SE.

A MSE based on learning from query logs by using prediction of user requirements was proposed in [13]. Query similarity function was used to measure the similarity of the web page with respect to the given query. Authors used 7 queries and 5 functions to test the query. These functions were named as:

(i) keyword similarity ( $\text{sim}_{\text{keyword}}$ ) (ii) similarity using documents clicks ( $\text{sim}_{\text{click}}$ ) (iii) similarity using both keyword and document clicks ( $\text{sim}_{\text{combined}}$ ) (iv) query clustering and (v) rank updater. Similarity based on query keywords were used for similarity calculation and clustering the results. While calculating similarity authors considered clicked URLs and Bipartite graph of query log also. Finally the combination of proposed similarity measure and clustering algorithm was used to cluster the queries.

### 3. PROBLEM FORMULATION

The major problems of MSEs [1], [14], [15], [16], [17] are discussed below:

- As almost every MSE just receive the results from multiple search engines & does nothing for presentation of these results. They present the results based on first come first serve basis. So a need arise for the MSE that can provide better presentation of results to user.
- A MSE proposed by [17] have shown good results but calculation in relevancy calculation may take more time which limits the significance of the returned results
- Some available literatures present the results using positional ranking and count function but such type of ranking fails to provide relevant results to user.
- Some MSEs decides number of clusters to be generated in advance. But no method was developed if numbers of returned URLs are more than the expectation.
- Several MSEs like Clusty generate the clusters and named them based on the

maximum number of query keyword occurred in the document. But when user going to search these cluster then they found that clusters have nothing as per their name. So there is some problem in deciding the name of clusters also.

- f) According to M. Kobayashi et al. most of the times the user's fires topic specific queries but remain unsatisfied from the results returned by the MSE. All this happens due to the low quality relevancy algorithm used by the MSEs.

#### 4. PROPOSED ARCHITECTURE

Proposed architecture is drawn in Fig. 1, which uses both ranking and clustering for organizing and presenting web searched results. The description of this architecture is organized in modules with description:

**i) Consumer Interface (CI):** CI is the way of interaction to the outer world from where user gives his searched query and gets the desired results.

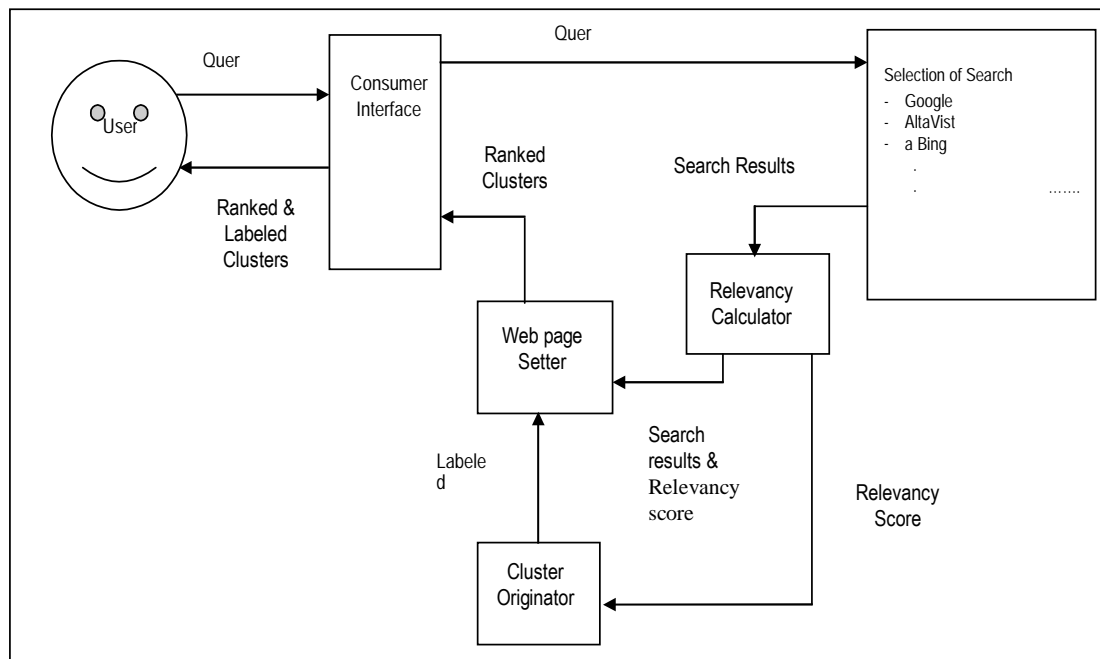


Fig. 1: Proposed MSE Architecture

**ii) Relevancy Calculator (RC):** RC assigns some relevancy value to each incoming URL of SEs. As there are many methods like VSM, OSM, CDR etc to calculate the relevancy of a returned webpage. But the relevancy calculation in these methods are seems to be complex which can be reduce to some extent. As Naresh Kumar and R.

Nath [17] uses VSM in their literature. They use number of terms, length of document and some logarithmic calculations but same work can be done by calculating the number of times query terms occurred in the document. Almost same results will be returned by this method which may result in reduction of time and space complexity also.

**iii) Cluster Originator (CO):** CO creates the desired number of clusters. The rest process of cluster generation will remain same as Cluster Generation module as explained in [17].

**iv) Web Page Setter (WPS):** The main task of WPS is to remove the replica of webpages from the results and sending the ranked web pages to the related cluster. The remaining process of WPS will remain same as WPA explained in [17].

**v) Selection of Search Engine (SSE):** SSE provides the number of search engines to be used for searching the user query. User can select any number of listed search engines. But selection or the use of more number of search engines may affect the performance of the MSE. So before selection of search engine to be used a user must know the domain expertization of the search engine which may help the user to get the result effectively and efficiently. It also fastens the processing of MSE.

## 5. CONCLUSION:

An effective MSE architecture for better presentation of results has been proposed in this paper. The proposed architecture discussed the problems related with the existing (like Clusty) as well as MSEs available in the literatures. Further a good suggestion for relevancy calculation is also proposed in this paper which may result in reduction of both types of complexity i.e. time as well as space complexity. The proposed ideas of relevancy calculation will also help in detection of more relevant results and improvement in speed of the overall system.

## 6. FUTURE WORK

The author is currently working on development of the proposed MSE architecture. Moreover author is also investigating some other issues like reduction of load on the network while using multiple SE with respect to MSE, improvement in presentation of results, naming convention of clustering etc. which are needs to be improved much more. The same will be shaped with completion in near future. The results and conclusion will be compared with the other techniques used by the existing MSEs.

## REFERENCES

- [1] P. Biraj and D. B. Patel, "Ranking Algorithm for Meta Search Engine", in International Journal of Advanced Engineering Research and Studies, E-ISSN2249-8974, Vol. II, Issue I, pp. 39-40, Oct.-Dec. 2012.
- [2] Rajender Nath 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, pp. 11-14 Jan - June 2013.

- [3] R. Nath and N. Kumar, "A Novel Parallel Domain Focused Crawler for Reduction in Load on the Network" published in International Journal of Computational Engineering Research, ISSN 2250-3005, Vol. 2, Issue. 7, pp. 77-84, Nov. 2012.
- [4] N. Kumar, R. Nath and P. Kherwa, "An Automated Framework based on TLS to choose Best Search Engine in a particular Domain" in International Journal of Computer Applications", ISSN online: 1741-5047, pp. 42-48, 2013.
- [5] Y. Lu, M. Weiyi, S. Liangcai, Y. Clement and L. K. Lup," Evaluation of Result Merging Strategies for Metasearch Engines", 6<sup>th</sup> international conference on web information engineering, pp. 53-66, 2005.
- [6] C. W. Tsai et. al., "A Document Clustering Approach for Search Engines", in IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan, pp.1050-1055, October 2006.
- [7] S. Zhu, X. Deng, K. Chen, and W. Zheng, "Using online relevance feedback to build effective personalized Metasearch engine", In proc of IEEE, 2<sup>nd</sup> international conference on Web information systems Engineering (WISE'01), Kyoto, Japan, Vol.1, pp. 262 – 268, 2002.
- [8] Z.Wu, W. Meng, C. Yu and Z. Li," Towards a highly scalable and effective meta search engine", In proc. of 10<sup>th</sup> international conference on World Wide Web, Hong Kong, pp. 386-395, 2001.
- [9] Y. Fu and D. W. Jin," An Implemented Rank Merging Algorithm for Meta Search Engine", in International Conference on Research Challenges in Computer Science, pp. 191 – 193, 2009.
- [10] A. Gulli and A. Signorini, "Building an Open Source Meta Search Engine", in proceeding of WWW 05 Special interest track and posters of the 14<sup>th</sup> international conference on world Wide Web, ISSN 1-59593-051-5, pp. 1004-1005, 2005.
- [11] Y.Y. Fu and W. J. dong, "An Implemented Rank Merging Algorithm for Meta Search Engine", in International Conference on Research Challenges in Computer Science, pp. 191 – 193, 2009.
- [12] D. Minnie and S. Srinivasan, "Meta Search Engine with an intelligent Interface for Information Reterieval on Multiple Domains", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.4, pp. 37-45, 2011.
- [13] G. Geetanjli, O. Ashish and O. Mamta, "Query Recommendation Approach For Searching Database Using Search Engine", in International Journal of Research in Engineering & Applied Sciences, ISSN: 2249-3905, Volume 3, Issue 3pp. 146-154, 2013.
- [14] Z. Oren and E. Oren," Grouper: a dynamic clustering interface to Web search results", in Proceeding of WWW '99 Proceedings of the eighth international conference on World Wide Web, Elsevier North-Holland, Inc. New York, NY, USA, Volume 31 Issue 11-16, Pages 1361-1374, May 17, 1999.
- [15] C. Douglass, K. R. Devid, P. Jan and T. John, "Scatter/Gather: a cluster-based approach to browsing large document collections", in Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), pp 318-329, 1992.
- [16] K. Mei and T. Koichi,"Information retrieval on the web," in ACM Computing Surveys,

vol. 32, no. 2, pp. 144-173, 2000.

- [17] N. Kumar and R. Nath, “A Meta Search Engine Approach for Organizing Web Search Results using Ranking and Clustering”, in International Journal of Computer (IJC), volume 10, No 1, pp. 1-7, ISSN 2307-4531, Oct. - 2013.