https://doi.org/10.26438/ijcse/v6si1.5559

IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Special Issue-1, Feb 2018 E-ISSN: 2347-2693

Architecture for Personalized Meta Search Engine

N. A. Borkar[1*], S. V. Kulkarni[2]

[1*]Department of Computer Engineering, Dhole Patil College Of Engineering, Pune, India
[2] Department of Computer Sci. & Engg., Walchand College of Engineering, Sangli, India

[*]*Corresponding Author:* borkarnutan9@gmail.com, *Tel.: +91-9158461475*

Available online at: www.ijcseonline.org

***Abstract*—** Information available on the web is growing rapidly. A major problem in web search is that the interactions between the users and search engines are limited by the factors like unknown capabilities of search engines adopted, and ill-constructed query by the user. Hence the user has to repeatedly apply the several queries till he reaches the pages of most interest.

Any search engine can give its best performance if well-constructed and detailed queries are used. As a result, the users tend to submit shorter/ insufficient/ ambiguous queries yielding unwanted search lists. In order to return highly relevant results to the users, search engines must be able to profile the users' interests and personalize the search results according to the users' profiles. This paper discusses the need and specific requirements of personalized search engine, its architecture, the prototype model developed and the results obtained. Also sample sessions performed on the designed model have been given for selected user profile.

***Keywords*—** Web Search Engines, Personalized Web Searching, Meta Search Engines.

I. INTRODUCTION

Search Engines have played important role in web Information retrieval (IR) systems, which uses a keyword index for the given input search query and in response returns a ranked list of resulting text in the form of short information objects called snippets. The ordering of the resulting snippets depends on how the web pages are ranked by the page ranking algorithm used by the specific search engine. This certainly differs from one search engine to another. Therefore we find that even most popular search engine like Google, Yahoo, and Bing tend to give different ranking for the same query or keywords applied. The figure 1 shows the query process, in which user sends a query to the searching, the query will go to the document database, then the page ranking will be done and the results will posted by the search engine.

Many search engines these days are providing auto suggestions to get a pin pointed query tag so that pre-cached results can be returned at the earliest. Search engine would give its best performance if well-constructed and detailed queries are applied, otherwise users will get lot of unwanted snippets in response to shorter/ insufficient/ ambiguous queries inputted by them.

One of the tricks used by the user is searching with many search engines for the same queries or tag-words. Meta-Search Engines are powerful tools that search multiple search engines simultaneously. Meta-Search Engine is a search tool that sends user requests to several other search engines and aggregates the results into a single list or displays them according to their source (on specific page and position based on rank).

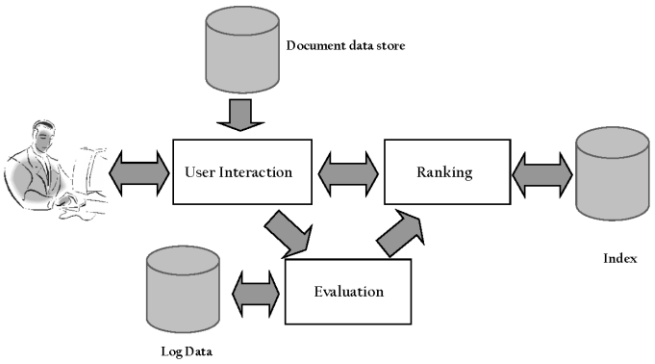

Figure 1. Query Processing in Search Engines

Meta-Search Engine (or aggregator) enables users to enter search criteria once and access several search engines simultaneously. Meta-Search Engine operate on the premise that the web is too large for any one search engine to index it all and that more comprehensive search results can be obtained by aggregating the results from several search engines. However, meta-search engines often do not search the best search engines because of ambiguities in the query

string formation for different search engines due to non-compatibility of query syntax.

The web searching has faced challenges like vastness, vagueness, uncertainty, inconsistency, and deceit. This is mainly due to the reason that search engine parser uses a specific page ranking algorithm designed to meet specific objectives. There for the passer algorithm has to address all of these issues in order to deliver appropriate web lists.

Another problem in web searching is lack of personalization since the motives behind searching vary from user to user though the same tag-words are getting used. Which search results are relevant varies is left to personal judgment of the user. Therefore in order to combine benefits accessing simultaneously multiple search engines and also to configure the personal style of user, we are presenting architecture of personalized-meta-search-engine.

In this paper, we address the architectural modeling for 'personality-profile driven context sensitive query-result aggregation' to help in converging the search list volume by redundancy elimination while yielding the most significant web pages. The paper organized as follows, Section I contains the introduction of issues, need and specific requirements in web searching, Section II discusses the literature referred related to personalization as well as aggregation methods in web searching, Section III contain related work done in designing the architecture of Personalized Meta Search Engine we named **MindSearch**, section V describes results and discussion, and Section VI concludes research work with future directions.

II. RELATED WORK

A Meta-search engine (or aggregator) is a search tool that uses another search engine's data to produce their own results from the Internet. Meta-search engines take input from a user and simultaneously send out queries to third party search engines for results. In order add user perspective into the search personal style and semantics that is based on mental context behind search, improvements are needed in meta search engine architecture to come out with personality or archetype or even location driven searching on web.

The personalization in query formulation based on the location for mobile users proposed by K W T Leung, and others [1]. They proposed a realistic design for PMSE server by adopting the meta-search approach which relies on one of the commercial search engines, like Google, Gigablast, Lycos or Bing. In their framework, the client is responsible for receiving the user's requests, submitting the requests to the PMSE server, displaying the returned results, and collecting his/her click through in order to derive his/her personal preferences.

Towards the personalization of search engine queries many authors have contributed by suggesting the data clustering techniques. A way to add such enhancement to personal search has been proposed by Prakasha, S. and et al [2], highlighting importance of design for information retrieval using structured intelligent search engine by employing query clustering technique and the concept of semantic web. While Annadurai A. [3] discussed improvisation using suffix tree clustering. A different approach has been proposed by Leung, W. Ng, and Lee [4] for concept-based clustering.

J. Teevan, S.T. Dumais, and E. Horvitz. [4], presented possibility to provide effective and efficient personalized Web search using a rich and automatically derived user profile, examined new algorithmic and user interface approaches that can improve personalized search.

III. METHODOLOGY & ARCHITECTURE OF MINDSEARCH

Advangle (http://advangle.com) which is a kind of search aggregator, that provides a simple and convenient builder of complex web-search queries for Google or Bing search engines, allows user to quickly build a query with multiple parameters and presents a combined list eliminating duplicate links. We have improvised the aggregation techniques of Advangle in terms of personalization by user profiling, feedback based query optimizer and presentation of results in multiple views to make it true personalized-meta-search-engine as shown in figure 2.

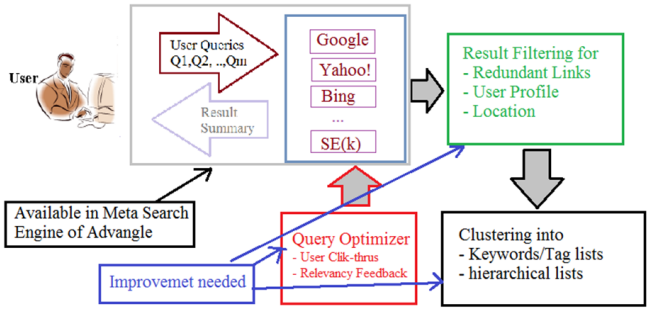

Figure 2. Improvement needed for Redundancy, Clustering and Query Optimization

It was also required to create a semantic profile of the user for monitoring and analyzing the users search history. The search results generated then utilized in amalgamation of varied techniques including clustering, re-ranking and semantics perceived from user profiles to enhance the performance of the web search engine.

The architecture of the **MindSearch** Framework includes the six main components:

1. **User Interface:** Allowing login of the user, storing the personal details like educational background, sex, age, personal web-pages, social network links, etc. which can be used to access the personality parameters. Also this facility is used to recognize and store the specific queries of the user so that user simply selects the previously applied queries as and when required.

2. **Query Parser:** The query parser transforms the query elements into token list. As different Search engines use different syntax, the SE query builder need to hand specific keywords, labels and tags. This processing is required on the client side as the selection of typical search engine is done by the user online.

3. **Result Aggregator:** The search result list is available in the form of snippets that include title, home-url, short-description and the file-type indicator is aggregated by collecting all the results for the specified query as-it-is.

4. **Redundancy Filters:** The framework require the pre-filters for removing duplicate snippets and make available only the unique links. The filter checks if the home-url is same or not as the same link may be captured with different title and short description by different search engines.

5. **Hierarchical Classifier:** The hierarchical classifier uses functional components like (a) Stop word remover so that exact keywords vector per snippet is created, (b) Word-frequency analyzer to count the most relevant words appearing in the snippet list,. (c). Hierarchical list builder which categorizes the list in two modes: **Unsupervised classification** that does not require user context & **supervised classification** that takes the user context into account.

6. **Tree Builder**: The reduced search list is presented in the form of expandable lists based on the classified results for specific query applied by the user.

Figure 3 shows the outlook of homepage of MindSearch PMSE while figure 4 displays the screen shot once the user is logged in and fires a query with preselected parameters.

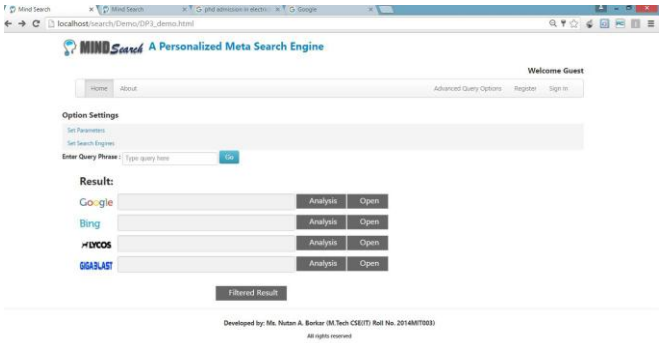

Figure 3. MindSearch Homepage View

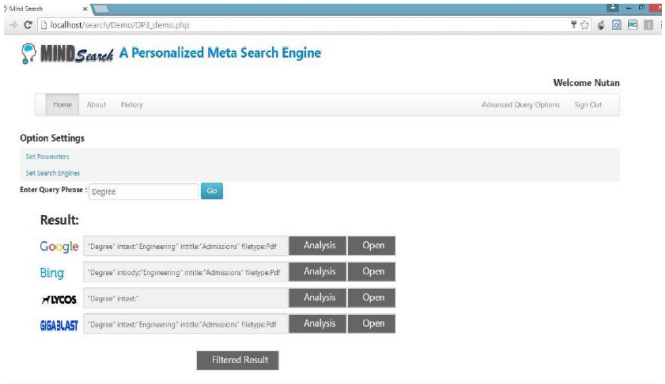

Figure 4. MindSearch View (post login)

Figure 5 show how the flow of the overall framework is interacts with the user. Firstly when the user is using the meta-search engine at that time if user us new then he can use the MindSearch directly or by registering his personal information. Once he register he can log into his account with user-id and password, or user is old user he can directly login. If the user is old user then he can enter the new query or he can see his old queries which he has saved in the account. And user can get the result.

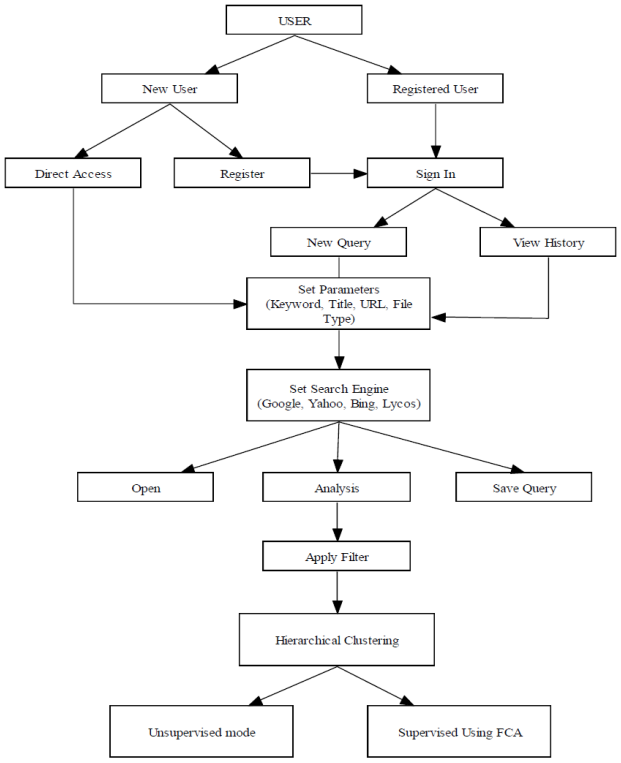

Figure 5. MindSearch Interaction Flow

IV. RESULTS AND DISCUSSION

The applicability of MindSearch was tested running various queries and aggregation results for top 20/30/50 results and then compared against the human observer (domain expert).

Table 1 shows the summary of **'Single Query term with N-Snippets per Search Engine'** from selected search engine and analyzed for the effectiveness of filtering and relevancy.

Table 1. Results of Single Query with N-Snippets per Search Engine

Analysis of Single query (Human Observer)

Query Parameters	Search Engines Used			Post Redundancy Filtering #	Post Relevancy Filtering #	% Reduction
	Google	Bing	Gigablast			
Keyword contains= big data; Title contains= big data analytics; Query Phrase= book	20	20	20	21	17	72.00%
	30	30	30	24	18	80.00%
	50	50	50	30	20	87.00%

Analysis of Single query (by MindSearch MSE)

Query Parameters	Search Engines Used			Post Redundancy Filtering #	Post Relevancy Filtering #	% Reduction
	Google	Bing	Gigablast			
Keyword contains= big data; Title contains= big data analytics; Query Phrase= book	20	20	20	21	12	80.00%
	30	30	30	24	13	86.00%
	50	50	50	30	16	89.40%

The result show that almost 60% percent snippets are irrelevant as these are ranked beyond query keywords & do not match the context of search. Also as the number of pages of search engine increased we get more accurate results where the accuracy increases gradually.

To test the effectiveness the results are analyzed first manually and then compared against the results produced by the framework. It is found that both the results are almost close and with error not more than 3%.

Another test was carried out for "**Differentiated Queries on Same Context**", e.g., when user wants to search books on "Big Data Analytics", he is likely to apply the query in different styles; this is a case of differentiated query phrase list. The Table-2 shows the effect varying the syntactical combinations of the specific query keeping the same context of book search on "Big Data Analytics".

Table 2. Differentiated Queries on Same Context

Analysis of Differentiated queries on same context (Human Observer)

Query Applied	Search Engines Used			Post Redundancy Filtering #	Post Relevancy Filtering #	% Reduction
	Google	Bing	Gigablast			
Keyword contains= big data; Title contains= big data analytics; Query Phrase= book	30	30	30	24	18	80.00%
Title contains= big data analytics; Query Phrase= book	30	30	30	20	12	86.00%
Query Phrase= big data analytics book	30	30	30	19	14	84.00%

Analysis of Differentiated queries on same context (by MindSearch MSE)

Query Applied	Search Engines Used			Post Redundancy Filtering #	Post Relevancy Filtering #	% Reduction
	Google	Bing	Gigablast			
Keyword contains= big data; Title contains= big data analytics; Query Phrase= book	30	30	30	24	18	80.00%
Title contains= big data analytics; Query Phrase= book	30	30	30	20	12	86.00%
Query Phrase= big data analytics book	30	30	30	19	14	84.00%

The results produced by manual observation as well as by the framework are almost similar within (+/- 5%) with redundancy of 80% to 90%. Thus using this framework the user gets most relevant results even with differentiated query phrases.

V. CONCLUSION and Future Scope

The architecture of MindSearch is provides a better performance against aggregator engines like "Advangle". The initial version achieved the following -

(1) Aggregation with large set of popular search engines by addition of Yahoo, Lycos and Gigablast Search engine with the immediate visualization of the results by the user.

(2) The redundancy in the final list by applying duplicate result elimination and adopting most relevant results from the aggregated search lists. The reduction in the list in most cases gives 80 to 90% of elimination giving only most relevant web page snippet references.

(3) Personalization helps in reducing the results when aggregation by location in the profile as well as location tracked by IP address is enabled.

The MindSearch architecture can further be improved for commercial version by including

(1) User Preferred Clustering Algorithms rather simply using default agglomerative hierarchical clustering (AHC) algorithm in the current version.

(2) Domain Specific Search Engines: The current framework uses free search engines like Google, Lycos, Bing, Gigablast which can be easily adopted with fewer code modifications. However, when user is very much interested in domain specific search engines like IEEE, Science Direct Paper Search, Amazon Search for variety of consumer products, the framework can give the scope for inclusion of domain aware parsers. For this kind of improvement the query formulator will need more enhancements.

.

REFERENCES

[1] K Wai-Ting Leung, D Lee, W Lee, *"PMSE: A Personalized Mobile Search Engine*", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, Issue: 4, pp.820-834, April 2013.

[2] S. Prakasha, H.Shashidhar, G.T. Raju, *"Structured Intelligent Search Engine for Effective Information Retrieval using Query Clustering Technique and Semantic Web",* International Conference on Contemporary Computing and Informatics (IC3I), 688 695, DOI: 10.1109/IC3I.2014.7019820.

[3] A Annadurai, *"Architecture of personalized web search engine using suffix tree clustering",* International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN 2011), pp. 604-608, 2011.

[4] K.W.-T. Leung, W. Ng, and D.L. Lee, "Personalized Concept-Based Clustering of Search Engine Queries," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 11, pp. 1505-1518, Nov. 2008.

[5] J. Teevan, S.T. Dumais, and E. Horvitz., "Personalizing Search via Automated Analysis of Interests and Activities. Proceedings of the 28th Annual International ACM SIGIR" Conference on Research and development in information retrieval (SIGIR'05), pages 449–456, 2005.

[6] Adah, S.; Bufi, C.; Temtanapat, Y., "Integrated Search Engine", @IEEE Knowledge and Data Engineering Exchange Workshop, 1997. Pages: 140 – 147.

[7] O. Zamir, O.Etzioni, "A Dynamic Clustering interface to Web search results," Computer Networks, Netherlands, Amsterdam, 31(11-16):1361-1374, 1999.

[8] M. Ilic, P. Spalevic, M. Veinovic, "Suffix Tree Clustering – Data mining algorithm", Twenty-Third International Electrotechnical and Computer Science Conference ERK'2014, Portorož, ISSN 1581-4572, pp. 15-18, September 22-24, 2014.

[9] K A Heller, Z Ghahramani. "Bayesian hierarchical clustering", Proceedings of the 22nd international conference on Machine learning, pp. 297-304, 2005.

[10] R.E. Ruviaro Christ, E. Talavera, C. Maciel, "Gaussian Hierarchical Bayesian Clustering Algorithm", ISDA 2007, pp. 133-13.

Authors Profile

Miss Nutan Borkar has pursed B.Tech in Computer Science & Engineering from SGGS College of Engineering, Nanded in 2013 and M.Tech in Information Technology from Walchand College of Engineering, Sangli in 2016.. She is currently working as Assistant Professor in Department of Computer Engineering at Dhole Patil College Of Engineering, Pune, India.

Dr S.V.Kulkarni pursed Bachelor of Electrical Engineering in 1984, Masters in Electrical Control Systems in 1991, Ph.D. in Electronics from Shivaji University Kolhapur, India. Also he is M.S.(Software Systems) from BITS, Pilani, India. His main research work focuses on Search Engine Architectures, Big Data Analytics, Data Mining, IoT and Inteligent Web Dynamics. He has 21 years of teaching experience and 12 years of Industrial Experience.

