IMSS: A Novel Approach to Design of Adaptive Search System Using Second Generation Big Data Analytics

Dheeraj Malhotra and O.P. Rishi

Abstract In this present era of Big Data, different search engine users have different information requirements at different intervals of time. Thus, search results should be adapted to user's requirements [1, 2]. In this research work, we propose a novel approach to adaptive web search augmented with capabilities of carrying out Big Data Analytics using second generation HDFS. Moreover, unlike conventional personalization techniques, the proposed approach does not require additional efforts from user such as reporting feedback/ratings etc. The proposed system can be implemented in the form of Intelligent Meta Search System (IMSS Tool) to overcome the problem of irrelevant web page retrieval faced by user of generic search engines. An extensive experimental evaluation shows that the average ranking precision of adaptive IMSS tool improves with trial runs when compared with a popular search engine.

Keywords Second generation HDFS · Personalized search · Big data search system · Meta search engine · Intelligent meta search system (IMSS) tool · Adaptive web search

1 Introduction

Adaptive search when supported by HDFS-Cloud framework leads to easy and efficient analysis of Big Data available on WWW to retrieve useful personalized page ranking patterns. Search engines are known to retrieve far larger information

D. Malhotra (✉) · O.P. Rishi
Department of CSI, University of Kota, Kota, Rajasthan 324005, India
e-mail: Dheerajmalhotra@ymail.com

O.P. Rishi
e-mail: Omprakashrishi@yahoo.com

N. Modi et al. (eds.), *Proceedings of International Conference on Communication and Networks*, Advances in Intelligent Systems and Computing 508,
DOI 10.1007/978-981-10-2750-5_20

but still no search engine can index more than about 16 % of index able web [3, 4]. The issue is not just only the volume but is also the relevancy with respect to user's information needs [1, 2]. When the same query is searched by different users, even a state of art search engine returns the same result, irrespective of the user submitting the query. For example, if a user is tech savvy and usually searches for laptop/mobiles then an incomplete query search like *Blackberry* should return documents related to *Blackberry mobiles* by intermediately expanding the query rather than returning the documents of some fruit. There are various types of conventional personalized search systems as discussed in literature. However these search systems fail to satisfy the user personalized requirements without having explicit ratings/feedback from user. Moreover such systems can't handle second generation Big Data as they not just require scalability, partial failure support etc. but also need to support multiple analytic methods on varied data types, as well as the ability to respond in near real time.

2 Contribution from the Study

To the best of our knowledge, this proposed research work is the first formal attempt to design and development of adaptive search system using intelligent big data analytics and is also deployable on cloud framework. Various contributions of the proposed approach may be summarized as follows:

- The user effort for providing explicit ratings/feedback in order to use personalized search system will no longer be required.
- The proposed system will overcome the limitations of traditional mining approaches to extract useful web search and page ranking patterns from Big Databases of Search engines by providing features like Scalability, Partial Failure Support etc.
- The proposed research work discusses the design of future ready intelligent search tool i.e. *IMSS* which can well satisfy the requirements of next generation Big Data Search System such as Real time response, support of multiple analytic engines.

3 System Design

The proposed system will follow modular approach as shown in Fig. 1. Here we first accept user search query and expand the same to intermediate query based on user's preferences obtained from his search history [5–7]. Proposed system will build user profile using user's long term and short term preferences derived from browsing history of n days ago and of current day of usage respectively. Meta key word recommender is used to derive Meta keywords of search from extracted

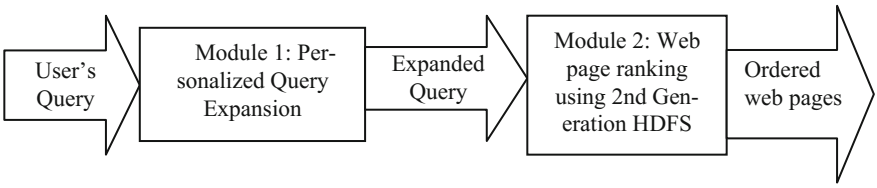

Fig. 1 Simplified design of proposed system

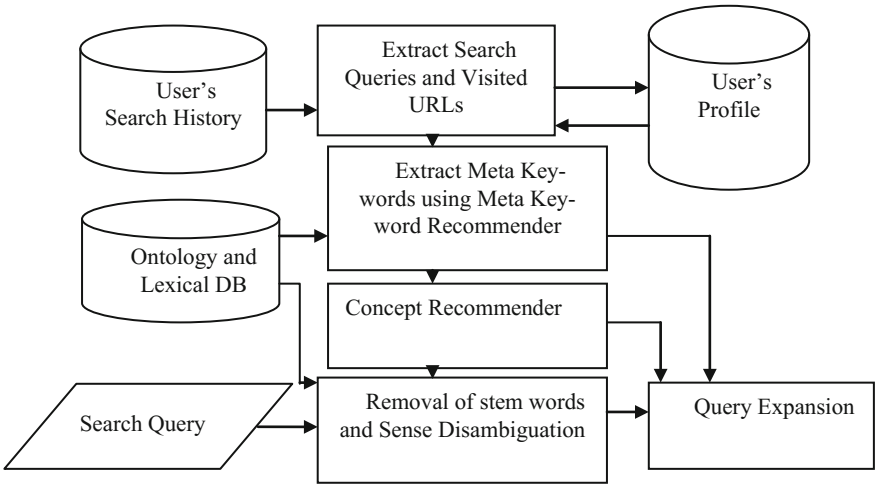

Fig. 2 Design of Module 1—personalized query expansion/modification

URLs. Similarly, Concept Recommender and Word Sense disambiguation processes are used for expanding user query into non ambiguous and more meaningful query as shown in Fig. 2. Module 2 is used for ranking of web pages obtained from backend search engines. HDFS Map() and Reduce() approach is used to calculate content relevancy vector; other relevancy vectors such as semantic relevancy vector (SRV) to determine the semantic closeness of user query with respect to web document under consideration, similarly Time Relevancy Vector is based on importance given by previous user of same web page. The detailed functionality of module 2 to determine weighted rank of candidate web page is shown in Fig. 3.

4 Second Generation HDFS and Map Reduce

There are two significant trends of Second Generation Big data Systems [2, 8] that are responsible for choosing second generation HDFS as a preferable deployment framework in proposed approach. (i) There is rapid growth in network bandwidth as

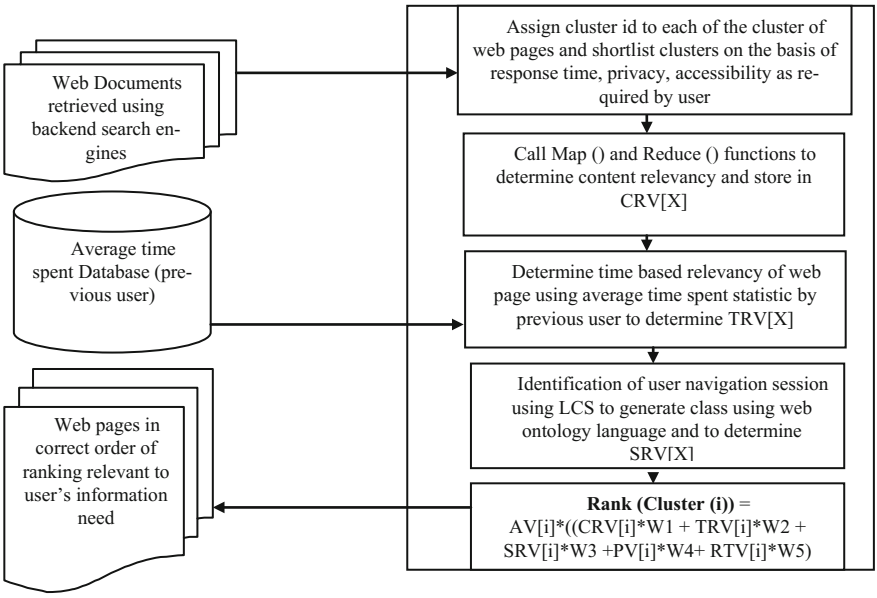

Fig. 3 Design of Module 2—web page ranking using HDFS based cloud framework

compared to hard drive bandwidth (ii) Development of In-memory computation models is urgently required to allow intermediate results to be kept in memory and hence reduces overhead of iterative analytics as suffered by conventional HDFS [9, 10].

HDFS is now adapted as long term store from which applications read their initial data and write their final results. The data layer is divided into sub layers for consistent storage and for intermediate objects separately to handle second generation of Big Data as shown in Fig. 4. In our proposed System, Map function will accept cluster ID as key and cluster log as second argument to tokenize each of web link entry in cluster log, obtained from back end search engine used by IMSS tool, to count individual occurrence of each of the keyword of search query. Extract () function is used to generate elements in list one at a time. Reduce function is coded to aggregate over all the occurrence of each keyword as provided by Map ()

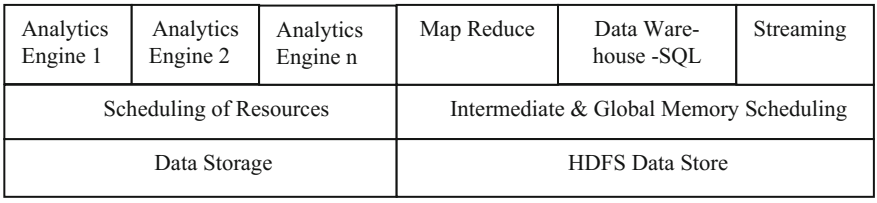

Fig. 4 HDFS deployment framework for second generation big data system

function [11] to determine keyword frequency in each of the web document and hence to determine the content relevancy vector. Map and Reduce code to be used by **Proposed System** is as follows:

```
Map (Int ID, String Log){
      List<String> X = tokenize (log)
        For each Token in X {// Token - Link extracted
                             //from back end search engine
Extract ((String) KWL, (Int) 1) // KWL – Keyword list
        }}
Reduce (String Token, List <Int> count)
              Int F = 0
             For each word in KWL {
                    F = F + 1
//F- Frequency count of each keyword
extract((string) token, (Int) F)}
```

5 Intelligent Meta Search System

In order to evaluate the proposed research design, *IMSS* tool using HDFS framework for analytics of second generation of Big Data is implemented using ASP. NET framework. The interface of IMSS tool is shown in Fig. 5. After Sign In, the inter-face of tool may allow user to select some or all of the four popular search engines like Google, Yahoo, ASK and Bing, for the purpose of intermediate web pages retrieval and search box allow user to specify search string. After clicking the Search button, tool will assign personalized rank to some of the top web links retrieved from back end search engines based on the calculation of various ranking vectors such as AV, SRV, CRV, TRV, RTV. The tool will return web links in the order of their ranking along with statistic of selected advanced search criterion. However *Take Me Fast* tab will not allow selecting any of the search criteria and will give result directly on the basis of user's history of browsing patterns stored in user's contextual database, which could be retrieved using his/her profile.

6 Comparative Precision Analyses—IMSS Tool V/S Google

In order to evaluate the effectiveness of our proposed approach, we recruited 10 human volunteers with age varied from 20 to 50 years with minimum of 5 years web search experience. 6 of them were males, 4 were females. They are asked to bring their personal laptops with installed IMSS tool followed by initial profile sign

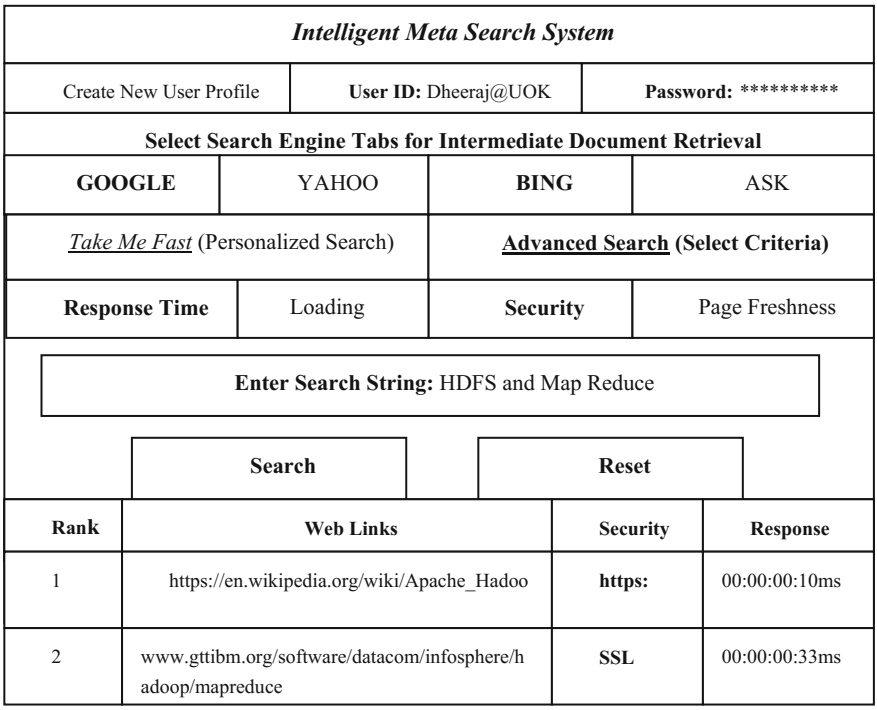

Fig. 5 Interface of IMSS tool

up process on tool, we followed following steps and asked volunteers to repeat the process for at least 4 trial runs one by one on Tool and Google:

1. In the first step, we asked volunteers to search an intentional incomplete query, for example a query like *Black Berry* rather than *Black Berry Mobiles* or *Black Berry Fruit.*
2. In the second step we asked volunteers to give points from 0(worst) to 5(best) to various precision parameters such as personalized page relevancy, page freshness, page size and response time to the top 10 links with respect to their shown rank in output of IMSS and Google.
3. After collecting data from each of the volunteer, we normalized the value of various precision parameters using expression:

$$Q_{\{ab\}} = (\mathrm{HIG}\,(P_{\{ab\}}) - P_{\{ab\}}) / (\mathrm{HIG}\,(P_{\{ab\}}) - \mathrm{LOW}\,(P_{\{ab\}}))$$

where, P_{ab} = Value of b_{th} Parameter of a_{th} web page; Q_{ab} = Normalized value of b_{th} Parameter of a_{th} web page; LOW, HIG = Lowest and Highest value of each of the parameter of precision.

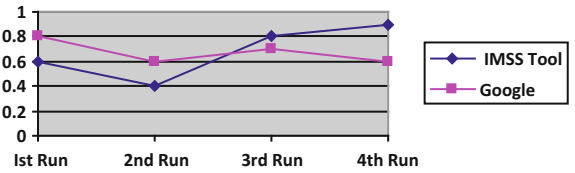

Fig. 6 Personalized precision comparison of IMSS tool with Google for Query "*BlackBerry*"

4. In the next step, we calculated the overall weighted precision of each web page retrieved by each volunteer as $N_a = \sum W_b \cdot Q_{ab}$, where, N_a = weighted precision of a_{th} web page; W_b = Weight assigned to b_{th} parameter by volunteer, usually $0 <= W_b <= 1$
5. Finally we determined overall precision by calculating average of all the weighted precisions as obtained from volunteers, Precision = AVG (N_a).

6.1 Observation

The graphical analysis in Fig. 6 shows that during first trial Run, precision of Google is reported as high; however with increase in number of trial runs, average precision of Tool improves slowly over Google. This is due to the fact that that Tool will build user profile and by employing personalized search can better satisfy the user for incomplete or ambiguous queries; On the other side, generic search engines try to interpret the query with all possible meanings without considering the preferences of user who searched for query and hence fails to achieve high value of personalized search precision.

7 Conclusion and Future Work

This research work present a HDFS based adaptive search framework for analytics of second generation of Big Data through implementation of IMSS Tool. The effectiveness of proposed approach is justified by experimental evaluation and comparison of personalized precision of IMSS tool over Google. The proposed approach can be applied to retail transactional or E Commerce website database as such transactional databases are also growing in the scale of Terabytes on daily basis and hence they require second generation Big data analytics system to mine useful customer buying patterns rather than conventional data mining techniques. The proposed system design can be enhanced by incorporating other advanced technologies such as Back Propagation Neural Networks, SVM etc. to further improve the precision of tool.

References

1. Wasid, M., Kant, V.: A Particle Swarm Approach to Collaborative Filtering based Recommender Systems through Fuzzy Features. In: Procedia Computer Science, IMCIP, Vol. 54, pp. 440–448, Science Direct, Elsevier, Bangalore, India, August 21–23 (2015).
2. Gebara, F., Hofstee, H., Nowka, K.: Second Generation Big Data Systems. pp. 36–41,Cover Feature Outlook, IEEE Computer Society (2015).
3. Shou, G., Bai, H., Chan, k., Chen, G.: Supporting privacy protection in personalized web search. In: IEEE transactions on knowledge and data engineering, Vol. 26, No 2, pp. 453–467. IEEE (2014).
4. Kuppusamy, K.S., Aghila, G.: CaSePer: An Efficient Model for Personalized Web Page Change Detection Based on Segmentation. Vol. 26, pp. 19–27, Journal of King Saud University, Elsevier (2013).
5. Verma, N., Malhotra, D., Malhotra, M., Singh, J.: E-commerce website ranking using semantic web mining and neural computing. In: International Conference on Advanced Computing Technologies and Applications, Elsevier Procedia Computer Science, Vol. 45, pp. 42–51. Elsevier, Mumbai, India, March 26–27 (2015).
6. Malhotra, D.: Intelligent Web Mining to Ameliorate Web Page Rank using Back Propagation Neural Network. In: 5th International Conference, Confluence: The Next generation information Technology Summit, pp. 77–81, IEEE Xplore, UP, India, September 25–26 (2014).
7. Malhotra, D., Verma, N.: An ingenious Pattern Matching Approach to Ameliorate Web Page Rank. Vol. 65, No 24, pp. 33–39, International Journal of Computer Applications, FCS, New York, USA (2013).
8. Khurana, A.: Bringing Big Data Systems to the Cloud. pp. 72–75, What's trending? Column, IEEE Computer Society (2014).
9. Tesai, C., Lai, C., Chao, H., Vasilakos, A.: Big Data Analytics: A Survey. 2:21, pp. 1–32, Journal of Big Data, SPRINGER (2015).
10. Singh, A., Velez, H.: Hierarchical Multi-Log Cloud-Based Search Engine. In: 8th IEEE International Conference on Complex, Intelligent and Software Intensive Systems, pp. 212–219. IEEE CPS, Birmingham, UK, July 2–4 (2014).
11. Son, J., Ryu, H., Yi, S., Chung, Y.: SSFile: A novel column-store for efficient data analysis in Hadoop-based distributed systems, Vol. 316, pp. 68–86. Elsevier Information Sciences, September 20 (2015).