Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 9 (2017), pp. 4609-4618

http://www.ripublication.com

Dynamic Navigation of Query Results Based on Hash Based indexing using Improved Distance Page Rank algorithm

R.M.Dharani Krishna, C.AnilKumar, K.Jeevan Pradeep and N.Papanna

Department of Computer Science and Engineering,
Sree Vidyanikethan Engineering College, Tirupati, Andhra Pradesh, India.

Abstract

World Wide Web as we know unlimited source of data, which contains list of internet pages and infinite links. During last ten years the size of web has grown, as millions of web pages are adding to web every day. So it is the most important source of information and more popular manner of communication. The main purpose of the Web Data Mining is to provide the hyper link structure for the internet pages.User entered queries on most of web sites, such as Cloud Bigtable (Cloud Bigtable is a publicly available version of Bigtable used by Google system) most of the time results in huge number of documents and hyperlinks, but only few results are related to user query, but they may not be present at top place. Web Page Ranking and classification of web queries, used in combination, to overcome the problem of non-relevant results for a given query. Web page results classification and most relevant information retrieval on a educational datasets is our proposed methodology. In our methodology, we propose a solution to this problem by categorization web queries dynamically using hash based indexing data structure and resulting web pages are ranked by using improved distance page rank algorithm. Using our proposed method, we reduced most of non-relevant results and most important results based on content and number of hyperlinks as top results for given query.

Keyword: Page Rank; Improved distance; concept hierarchies; dynamic navigation, cloud Big table

1. INTRODUCTION

In the past ten years the size of web increasing day to day, and it became the most important source for storing information and also used for communication [1]. The internet acting as a scaffold for interchanging various types of data in the form of educational data, data related to research, data related to images and videos, personal information, and types of software's and hardware's [9][3]. The main aim of web structure mining is to provide the hyper links between the web pages, and based number in links and out links, calculation of weight age of web page is determined [8]. Web Structure mining mainly performs classification of web pages and determination important web documents of all data the data present in web server [6][10]. Using web structure mining we calculate number of hyperlinks of each page and then by applying page rank algorithm, we provide most important results to user [12].

Most of search engines use the principle of document classification and page rank algorithm for sorting web pages [2]. It is considered as a model of patron conduct, where a user searches on connections [4] at arbitrary and not using admire towards content [5].The arbitrary surfer visits a web page with a specific probability which gets from this page Rank [6]. The chance that their regular surfer faucets on one connection is exclusively given through the amount of connections on that page [7][14]. That is the reason one's web page Rank is not definitely went on to a page it connections to, however is partitioned by the quantity of connections on the web page [15].

Most of methods groups internet queries primarily based on those three approved categories, the usage of k-medoids clustering algorithm and a feature-set fex illustration algorithm [8][10]. Every keyword in the search history log incorporates appearance along with person details, sessions, date and time of usage, keywords of query, and the kind of information the person is trying to retrieve from web[11][14]. Further, every keyword changed into enhanced with the question size, used by all most all search engines and web servers for retrieval for web documents for given query [12]. When the user visits any web page or document or performs ant transaction using during interaction session with search engines, that data can be categorized under any one the three types of navigational queries that is informational queries, transactional queries and navigational queries [6] [9].

Statically retrieval techniques mainly used to categorize web queries based totally on their purpose and number of hyperlinks [5][9]. This static categorization is then used to robotically classify new web queries thru genuine phrases comparable [7]. Despite, the technique is simply too prohibitive as it suitable for most related items [10]. So that we can address this problem, using classification strategies that use

mainly Vector Space model and Bayes classifiers [8], using a Vector space model received good outcomes compared to the informational retrieval using Bayes classifiers even though it used for retrieval of information with different intension [13][15]. Experimental results provide phrase-matching functions turn out to be key to understand useful web queries, but the performance it is poor in case of navigational queries [2].

Characterization or supervised training is one of the machine learning strategies for gathering related information from the given web servers data [1] [5]. This information mainly consist of an organization of marked cases which are sets consists of an information requires and a wanted most important results [2][6]. Though group learning alludes to a gathering of characterization strategies that take in an objective work via preparing various single classifier and consolidating their expectations [3][15]. The standard is that a panel choice, with individual expectations consolidated suitably, ought to have better general precision, all things considered, than any individual advisory group part [14].

The main objectives of our proposed method are mainly, to provide more relevant results to the given query fast and reduce the amount of time we spend searching for information, to reduce non relevant results for the given query, to overcome the limitations of existing navigation and Page Rank algorithm (used by Google) with Dynamic Navigation and Improved Distance Rank algorithm, and to reduce ads or popup windows or "phishing" for personal data.

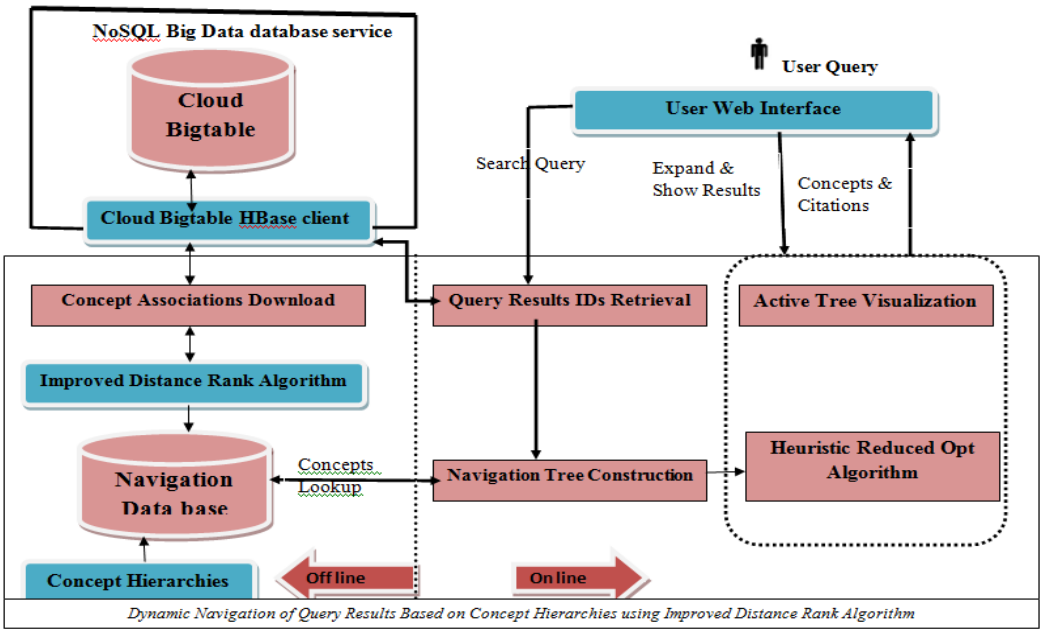

Figure 1: Architecture of our Proposed System

2. ARCHITECTURE OF THE PROPOSED SYSTEM

In this system, we proposed dynamic navigation of web queries based on Hash based indexing and Improved Distance Page Rank Algorithm. Dynamic navigation of queries can be performed by using hash based indexing and the algorithm for improved distance page rank is based on calculation of the distance between the authoritative pages and hub pages with good authoritative score and hub score and with the user can reach authoritative page from hub page with minimum number of clicks and can reach a hub page from authoritative page with minimum number of clicks. A good authoritative page for a given query is pointed by many good hub pages. A good hub page for a given query is pointing to many good authoritative pages.

Naturally, the distance between two hub and authoritative pages is the weight of shortest distance between I and j denoted dist ij which mainly based on number of in links and out links for i and j. Distance Page Rank algorithm recursively calculates the score of a web page and convert the value to in the scale for zero to one. Here, we use distance rank vector, which mainly consists of all distances arranged in ascending order that is lower distance web pages are assigned higher ranks. The time complexity of our algorithm will be good, as it requires less number of iterations. The complete evaluation of this distance calculation will be implemented in improved distance page rank. By using hash based indexing for categorization of web queries reduces the time taken to retrieve relevant documents. This method retrieves web documents if keyword present in query exactly matches with keywords present in the hash table, which reduces most of the non-relevant results. This method can also be used in multi-dimensional information retrieval for given query.

In addition we use a novel search interface as shown in figure 1, by using which the user enter a search query, all the data is present on cloud's Big table. First we construct a navigation hierarchy using hash based indexing and results obtained from hash based indexing are given to improved distance page rank algorithm to arrange the results in descending order with higher rank pages as top results. We used Hash based indexing method to selectively reveal the best concepts related to each keyword present in the given query. Using our proposed method mainly minimizes the navigation cost and it reduces time required for navigation. So our algorithm provides efficient results even for multi dimensional data.

3. ALGORITHM FOR DYNAMIC NAVIGATION OF QUERIES BASED ON HASH BASED INDEXING USING IMPROVED DISTANCE PAGE RANK

Algorithm: Dynamic Navigation and improved distance rank **Input**: Query entered by user in novel search interface.

Output: Web pages with ranks in descending order

1. Read the query entered by user
2. List out the keywords present in the given query
3. For all keywords present in the search query construct hash based indexing dynamically
4. Search for keywords present the query using extended hashing
5. Number of keywords in taken as n
6. Matching for given keywords is done by using
7. $b \leftarrow h(k) \,\&\, (2\,n - 1)$
8. Return b buckets matched for the given keywords
9. Apply this procedure recursively until all documents for keywords present in query are retrieved
10. All documents are arranged in descending order of their ranks
11. Calculate the count of number of incoming links and outgoing links for every document
12. Calculate unique visit count of web page
13. Calculate the distance the web pages
14. Rank are assigned to web pages from highest to lowest with minimum distance between web pages, highest unique visit count of web page and highest in links.
15. Display the web pages with ranks in descending order.

4. EXPERIMENTAL EVALUATION OF ALGORITHM

Evaluation of our algorithm is based on mainly two factors; firstly numbers of relevant documents retrieved and retrieved documents are ranked using improved distance page rank algorithm. We compared our results with graph based web

query classification[2] and agent based weighted page rank algorithm[5]. In the below table 1 we are representing results our algorithm compared with existing algorithm.

Table 1: Number of relevant web pages retrieved by Graph Based and Hash Based Classifications

Sno	Query No	No of Keywords	Count of web pages retrieved in Graph based classification[2]	Count of web pages retrieved in our proposed method
1	Q1	6	1200	778
2	Q2	7	1500	1200
3	Q3	8	1578	1350
4	Q4	6	1879	960
5	Q5	6	2500	879
6	Q6	5	1798	676
7	Q7	4	2357	766
8	Q8	9	3500	1236
9	Q9	4	2787	673
10	Q10	6	3250	985

4.1 COMPARISON OF RELEVAT WEB PAGES RETRIEVAL

Performance of our algorithm shows that hash based indexing for classification of web queries reduces most of non relevant results and produces most relevant results for the given query. We compared our results with Agent based weighted page rank algorithm [2]. Our algorithm produces proficient results as shown in figure-2.

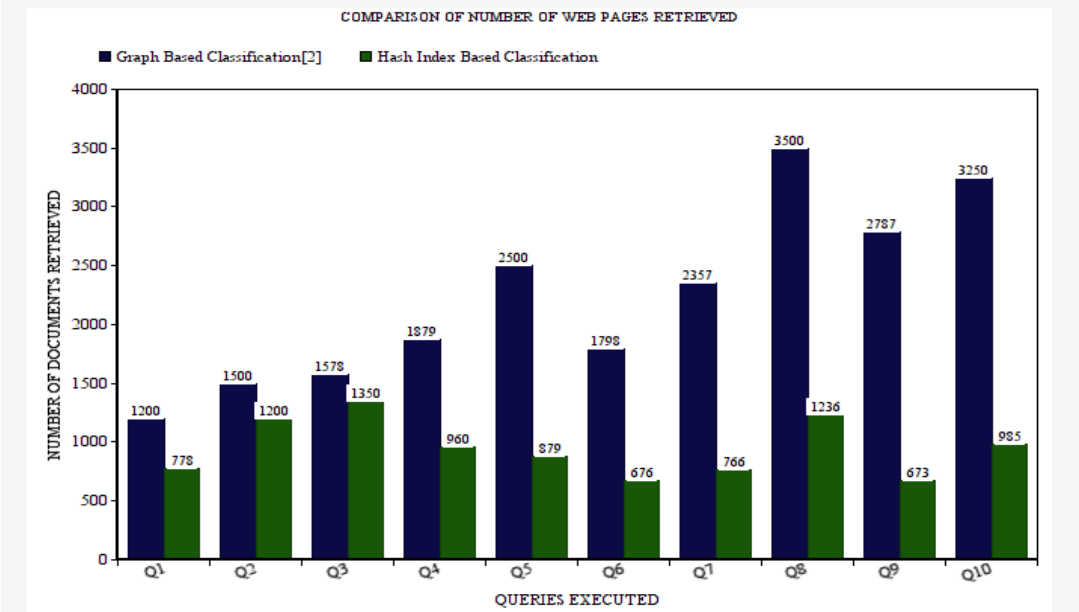

Figure 2: Number of web pages retrieved by Graph and Hash Based Classifications

4.2 COMPARISON OF PAGE RANKS OF WEB PAGES

After retrieval of relevant documents we calculated ranks of each web page using our improved distance page rank algorithm. We compared our results with Agent Based Weighted Page Rank Algorithm [5].In the below table 2, we are representing results our algorithm compared with existing algorithm Agent Based weighted Page Rank. Results produced by our algorithm are more relevant to given query and time taken to retrieval of results also less compared to existing algorithm.

Table 2: Ranks Calculated for web pages using Agent Based and Improved Page Rank Algorithms

SNO	URL of Web Page	Rank by Agent Based weighted Page Rank[5]	Proposed Method Improved
P1	http://sohacogroup.com.vn/index.html	567	237
P2	http://www.amicidelgiocodelponte.it/index.ht	658	325
P3	http://www.uniaoparaobem.com.br/index.html	657	728
P4	http://www.hotelmajore.it/ck.html	379	297
P5	http://v-montazar.com/index.html	958	560
P6	http://www.eca.edu.au/index.html	3576	2305
P7	http://www.speyerseminar.de/index.html	875	329
P8	http://www.jsigw.com/index.html	956	526
P9	http://www.leftoverpets.org/index.html	780	379
P10	http://www.bsc-md.de/index.html	1545	540

Performance of our algorithm shows that Improved Distance Page Rank Algorithm for ranking of web pages produces most relevant results for the given query in the top position in terms of both content and number of incoming and outgoing links. We compared our results with Agent based weighted page rank algorithm [2]. Our algorithm produces proficient results as shown in figure 3.

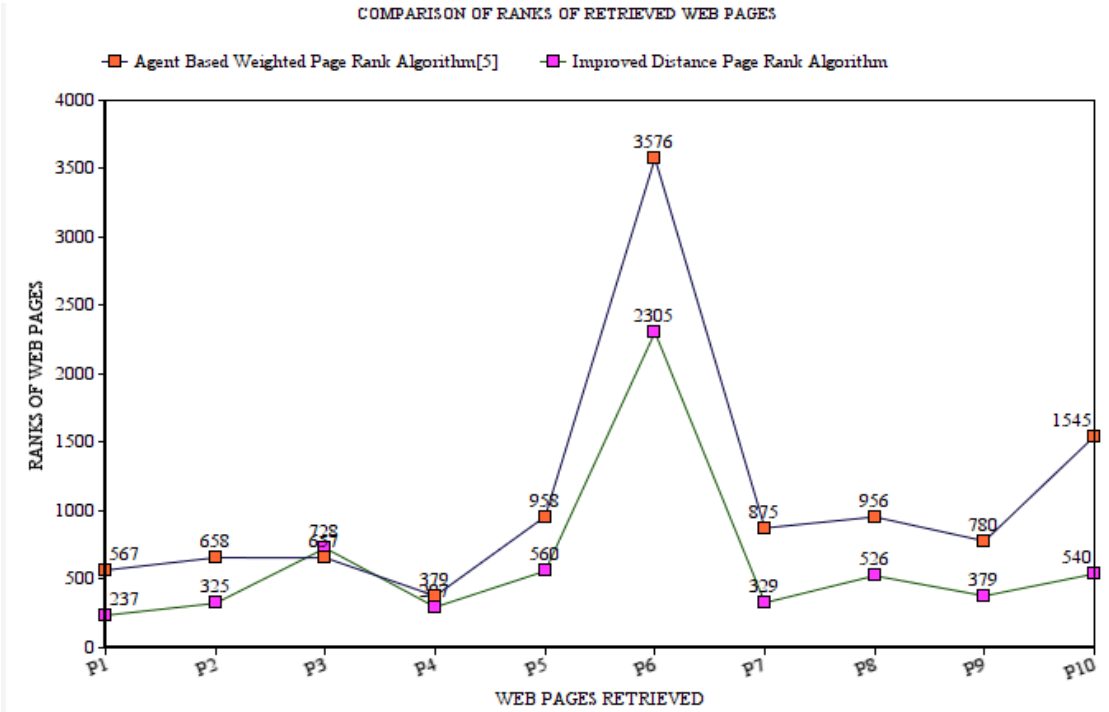

Figure 3: Ranks of web pages using Agent Based and Improved Distance Page Rank Algorithms

5. CONCLUSION

Retrieval of pertinent results for a given query is most complicated task; it depends on so many factors. In this paper we proposed an efficient method to reduce non relevant results for given query using dynamic navigation of queries using hash based indexing and improved distance page rank algorithm. This algorithm is based on extended hashing technique, unique visit count of web pages, distance between hubs and authorities of web pages. In this algorithm unique visit count gives us information about the web pages which are visited more number of times, so we will get more relevant web pages, in addition to this we are also considering the distances between hubs and authorities, if less distance between web pages then they can be reached in less amount of time. Our proposed method reduces most of

non relevant results for given query and retrieves more relevant documents as top results based on both content and hyperlinks. In this paper we developed dynamic navigation of queries using hash based indexing and Improved Distance Page Rank algorithm to retrieve more relevant web pages, by considering unique visit count, hub scores, authority score and distance between the web pages. Our algorithm works well with data base contains related data like medical data and education data bases. Development of efficient search algorithm for multi disciplinary data is a future scope of our work

REFERENCES

[1] Samir Amir, Hassan A¨ıt-Kaci1, "An efficient large scale reasoning method for the semantic web" Journal of Intelligent Information Systems, Vol 46, Issue 135, Nov 2016, 2-24.

[2] Chunwei Xia; Xin Wang, "Graph-Based Web Query Classification" IEEE Conference Publications,Pages: 241 – 244, Feb 2015.

[3] Nagappan, V.K, Dr. P. Elango, "Agent Based Weighted Page Ranking Algorithm for Web Content Information Retrieval", In proceedings of *International Conferenceon Computing and Communications Technologies,2015*

[4] Shailendra G. Pawar, Pratiksha Natani, "Effective Utilization of Page Ranking and HITS in significant Information Retrieval", in proceedings of International Conference for Convergence of Technology – 2014

[5] Alejandro Figueroa and Guenter Neumann. Exploiting user search sessions for the semantic categorization of question-like informational search queries. In International Joint Conference on Natural-Language Processing, pages 902–906, Jan 2013

[6] Areej Alasiry, Mark Levene, and Alexandra Poulovassilis. Extraction and evaluation of candidate named entities in search engine queries. In WISE, pages 483–496, Mar 2012.

[7] L Li, L Zhong, G Xu, and M Kitsuregawa, "A feature_free search query classification approach using semantic distance," Expert Systems with Application: An International Journal, vol. 39, no. 12, pp. 10739-10748, Feb 2012.

[8] Abhijith Kashyap, Vagelis Hristidis, Michalis Petropoulos, and Sotiria Tavoulari "Effective Navigation of Query Results Based on Concept Hierarchies" IEEE Transactions on Knowledge and Data Engineering, Vol 23, Issue 10, July 2011.

[9] H. Dubey and B. N. Roy, "An Improved Page Rank Algorithm Based on Optimized Normalization Technique," International Journal of Computer

Science and Information Techniques(IJCSIT), pp. 2183- 2188, 2011.

[10] X. Wang, T. Tao, J. T. Sun, A. Shakery and C. Zhai, "DirichletRank: Solving the Zero-One Gap Problem of"PageRank". ACM Transaction on Information Systems,Vol. 26, Issue 2, 2008.

[11] Mohammad Zareh Bidoki, Nasser Yazdani, "DistanceRank: An intelligent ranking algorithm for web pages" Internal journal of Information Proceesing and Management, 2007, 1-16.

[12] Ying Liu "Supervised HITS Algorithm for MEDLINE Citation Ranking " IEEE 7th International Symposiumon BioInformatics and BioEngineering, 14-17 Oct.2007, 1323 – 1327.

[13] Taher H. Haveliwala "Topic Sensitive Page Rank A Context-Sensitive Algorithm for Page Rank", IEEETransactions on Knowledge and Data Engineering, Vol 15, Issue 4, July 2003, 784-796.

[14] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Page Rank *Computer Networks and ISDN Systems*,30(1–7):107–117, 1998