IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-3 E-ISSN: 2347-2693

Dynamic Navigation of Query Results Using Biased Topic Sensitive Page Rank Algorithm

L. Lakshmi[1*] and P Bhakara Reddy[2]

[1]Department of Computer Science and Engineering, MLT Institute of Technology, Hyderabad, India
[2] Department of Computer Science and Engineering, Principal of MLT Institute of Technology, Hyderabad, India

www.ijcseonline.org

Received: Mar /09/2015 Revised: Mar/15/2015 Accepted: Mar/26/2015 Published: Mar/31/ 2015

Abstract*—** The major disadvantage of Page Rank is that it favors the older pages, because a new page, even a very good one will not have many links unless it is a part of an existing site. Page Rank is a global measure and is query independent. The Rank sinks problem occurs when in a network pages get in infinite link cycles. ***To improve the search results Topic-Sensitive Page Rank also referred to as TSPR is a context-sensitive ranking algorithm for web search developed by Taher Haveliwala. The disadvantage with topic sensitive page rank algorithm is it uses basis set is small that is it uses 16 top level categories. So we propose to improve topic sensitive page rank algorithm with best set of basis topics. Here we propose to use ***fine grained set of topics*** mainly categorized into four categories and sub categories and so on. Almost all paths end at maximum sixth level. This method results efficient results compared page rank and context sensitive topic sensitive page rank algorithms.

***Keywords*—** Page Rank; Context Sensitive Page Rank ;Biasing; Rank Sink; Link Cycles .

I. INTRODUCTION

Web Search Engine is a tool enabling document search, with respect to specified keywords, in the Web and returns a list of documents where the keywords were found. User interface is the part of Web Search Engine interacting with the users and allowing them to query and view query results[8]. There are so many web servers in the Internet and numerous web pages on each of them[4]. ıt is so important for any web search engine to rank the pages with the aim of providing more useful data, by listing the pages containing the data at higher places, to the searcher about the searched keyword or subject[1]. So to be able to provide desired ordering for the web pages:

1.1 A page ranking algorithm is the technique utilizing some valuable metrics about the web pages and ordering the pages accordingly[3]. Together with the development of the Internet and the popularity of World Wide Web, Web page ranking systems have drawn significant attention.Many Web Search Engines have been introduced until now, but still have difficulty in providing completely relevant answers to the general subject of queries. The main reason is not the lack of data but rather an excess of data. The “Page Rank” algorithm, proposed by founders of Google Sergey Brin and Lawrence Page, is one of the most common page ranking algorithms that is also currently used by the leading search engine Google[2]. In general the algorithm uses the linking (citation) info occurring among the pages as the core metric in ranking procedure. Existence of a link from page p1 to p2 may indicate that the author of is interested in page. It is executed at query time, not at indexing time, with the associated hit on performance that accompanies query-time processing[5]. Thus, the *hub(going)* and *authority(coming)* scores assigned to a page are query-specific. It is not commonly used by search engines. It computes two scores per document, hub and authority, as opposed to a single score of PageRank. It is processed on a small subset of 'relevant' documents, not all documents as was the case with PageRank.

1.2 Disadvantages of Page Rank Algorithm

- It is a static algorithm that, because of its cumulative scheme, popular pages tend to stay popular generally.
- Popularity of a site does not guarantee the desired information to the searcher so relevance factor also needs to be included[3].
- In Internet, available data is huge and the algorithm is not fast enough.
- It should support personalized search that personal specifications should be met by the search result[2].

1.3 Topic Sensitive Page Rank Algorithm

It is the Personalized version of Page Rank Instead of computing a single rank vector, we compute a set of rank vectors, **one for each (basis) topic** Uses the *Open Directory Project* as a source of representative basis topics (http://www.dmoz.org) or Yahoo!. Topic Sensitive Page Rank is Calculate in two steps, fully automatically :

- Pre-processing,
- Query-processing

Preprocessing step is calculated offline, just as with ordinary Page Rank

Query processing step is calculated online is if the query was issued by highlighting the term x in some Web page p, then x^1 consists of the terms in p. For ordinary queries not done in context, let x^1=x. Using a unigram language model, with parameters set to their maximum-likelihood estimates, we compute the class probabilities for each of the 16 top level ODP classes[2], conditioned on x^1. Let x_i^1be the ith term in the query (or query context) $x^{1.}$

II. Related Work

Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection[4]. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages. In the business world, structure mining can be quite useful in determining the connection between two or more business Web sites. The determined connection brings forth a useful tool for mapping competing companies through third party links such as resellers and customers.

The main two problems with web structure mining are, the first of these problems is irrelevant search results. The second of these problems is the inability to index the vast amount if information provided on the Web[12].

ODP (Open Directory Project) is metadata 4 million sites, 590,000 categories. It is Tree Structure, it contains categories as inner node and Pages are leaf node, high quality, representative. Using ODP Metadata to personalize Search in 4 billion vs. 4 million, Using ODP Metadata for personalized search[13]. Biasing possible in the ODP context by

Extend ODP classifications from its current 4 million to a 4 billion Web automatically by biasing.

2.1 Several topics from ODP selected by user, Personalized Search,

Send *Q* to a search Engine *S*(E.g., Google, ODP Search), *Res*=URLs returned by S, For i= 1 to size(*Res*)

Dist[i]=Distance(Res[i], Prof)

Resort *Res* based on *Dist*

2.2 Representation

Both user profile and URL(50% in Google directory) can be represented as a set of nodes in the directory tree

2.4 Distance (Profile, URL)

Minimum distance between the 2 set of nodes.

2.5 Naïve Distances

Minimum tree distance, Intra-topic links, Sub Sumer

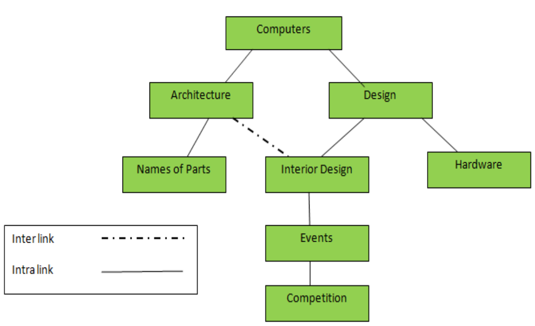
2.6 Complex Distance

The bigger the sub sumer's depth is, the more related are the nodes

$$s'(a,b) = ((1-\gamma)e^{-\alpha l_1} + \gamma . e^{-\alpha l_2}) . \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}$$

2.7 Combing with Google Page Rank

$$s''(a,b) = \delta \frac{1}{1+s'(a,b)} + (1-\delta).PageRank(b)$$

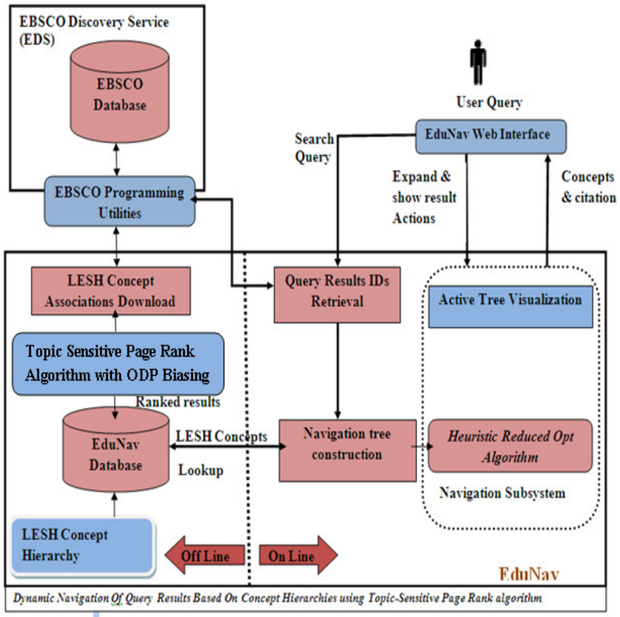


III. PROPOSED WORK

Dynamic navigation of query results uses multi level categorization and ranking are used to minimize the information overload and the navigation cost faced by the users. In this multi-level categorization, it executes and verifies at each step. Due to this verification process it produces only relevant data at each step. Here we use educational data base with dynamic navigation and applying context sensitive page rank algorithm with fine grained biased topics[8].

In this system, we present a topic-sensitive Page Rank algorithm, we pre compute the importance scores offline, as with ordinary Page Rank algorithm. However, we compute multiple importance scores for each page; we compute a set of scores of the importance of a page with respect to various topics[7]. At query time, these importance scores are combined based on the topics of the query to form a composite Page Rank score for those pages matching the query[6]. This score can be used in conjunction with other IR-based scoring schemes to produce a final rank for the result pages with respect to the query.

In addition we use EduNav novel search interface that enables the user to navigate large number of query results by organizing them using the LCSH concept hierarchy. First, the query results are organized into a navigation tree. At each node expansion step, EduNav reveals only a small subset of the concept nodes, selected such that the expected user navigation cost is minimized[9]. In contrast, previous works expand the hierarchy in a predefined static manner, without navigation cost modeling. We show that the problem of selecting the best concepts to reveal at each node expansion is NP-complete and propose an efficient heuristic as well as a feasible optimal algorithm for relatively small trees[4].



Dynamic Navigation Of Query Results Based On Concept Hierarchies using Topic-Sensitive Page Rank algorithm

The ODP biasing used in topic sensitive page rank is simple and it uses up to 16 categories topics. By analyzing the quality of the ODP biased Page Rank vectors that is Mean Average Precision (MAP). Our proposed ODP catalog, which we have recently gathered using the Heritrix 3 crawler. Around 100 biasing sub-categories were randomly chosen from four top level categories, namely Business, Computers, Recreation and Sports. The selection process was executed as follows: For each of the four top categories, three subcategories were randomly picked; then, for one of them, we again randomly took three subcategories and so on, until no deeper levels were available. Almost all paths ended at level 6 with level 1 being one of the ODP root categories. Finally, we computed Biased Topic Sensitive Page Rank vectors using the pages residing in each of these categories as biasing sets. We also selected six queries per category randomly. Commonly used query terms to some specific keywords of interest. Whenever such a query resulted in less than one hundred results within our index, we replaced it by another one, randomly selected as well. Nevertheless, in most cases we obtained several thousands of results per query. Note that these queries are implicitly focused on each given ODP topic, and thus they should have resulted in rather similar outputs for Non biased and Biased Page Rank.

It is important to note that biasing Topic Sensitive Page Rank using ODP is highly useful in many applications. To name but a few, Combating Web Spams (Trust Rank), Ranking Images, it can be employed for Personalized Web Search (i.e., bias on user's topics of interest), Faceted Search (i.e., promote the selected facet by biasing), Automatic Extension of the ODP (i.e., derive new qualitative pages to add into each category).

IV. CONCLUSION

The web as we all know is an infinite source of information which includes massive collection of web pages and countless hyperlinks. From the massive collection of web pages we need to extract useful information, relevant information and we need reduce numbers citations. For this we proposed a dynamic navigation instead of static navigation and we want use topic sensitive page rank algorithm to get effective results with deeper levels of ODP biasing.

V. FUTURE WORK

The web as we all know is an infinite source of information which includes massive collection of web pages and countless hyperlinks. From the massive collection of web pages we need to extract useful information, relevant information and we need reduce numbers citations. For this we proposed a dynamic navigation instead of static navigation and we want use topic sensitive page rank algorithm to get effective results with deeper levels of ODP biasing.

REFERENCES

[1] shaojie qiao , tianrui li, hong li, yan zhu, jing peng, jiangtao qiu "simrank: a page rank approach based on similarity measure", (ıntelligent systems and knowledge engineering (ISKE), 2010 international conference)

[2] fabrizio lamberti, andrea sanna, and claudio demartini "a relation based page rank algorithm for semantic web search engines", (IEEE Transactions on knowledge and data engineering, vol. 21, no. 1, january 2009)

[3] junghoo cho sourashis roy robert e. adams "page quality: in search of an unbiased web ranking", (SIGMOD 2005 june 14-16, 2005, baltimore, maryland, usa.)

[4] arasu, j. cho, h. garcia-molina, a. paepcke, s. raghavan: "searching the web", ACM transactions on internet technology, vol 1/1, aug. 2001 chapter 5 (ranking and link analysis)

[5] S. BRIN, l. page: "the anatomy of a large-scale hyper textual web search engine", www 1998 chapter 2 and 4.5.1

[6] border, kumar, maghoul, raghavan, rajagopalan, stata, tomkins, wiener: "graph structure in the web", www 2000

[7] page, brin, motwani, winograd: "the pagerank citation ranking: bringing order to the web", stanford technical report haveliwala: "topic-sensitive pagerank", www 2002

[8] L. Lakshmi*, Dr. P. Bhaskara Reddy, Dr. C. Shoba Bindhu "Dynamic Navigation of Query Results Based on Concept Hierarchies using Topic-Sensitive Page Rank Algorithm " Dec 2014

[9] Utility Analysis for Topically Biased PageRank Christian Kohlschütter. Paul-Alexandru Chirita,deWolfgang Nejdl.

[10] W. Xing and Ali Ghorbani, "Weighted PageRank Algorithm", *Proc. Of the Second Annual Conference on Communication Networks and Services Research (CNSR* '04), *IEEE*, 2004.

[11] Taher H. Haveliwala. "Topic-Sensitive Page Rank: A Context-Sensitive Ranking Algorithm for Web Search". IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No4, July/August 2003, 784-796.

[12] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.

[13] N. Duhan, A. K. Sharma and K. K. Bhatia, "Page Ranking Algorithms: A Survey", *Proceedings of the IEEE International Conference on Advance Computing*, 2009.

[14] M. G. da Gomes Jr. and Z.Gong, "Web Structure Mining: An Introduction", *Proceedings of the IEEE International Conference on Information Acquisition*, 2005

Authors Profile

L.Lakshmi, I completed my M.Tech in Computer at Jawaharlal Nehru Technological University Hyderabad. I have 9 Yrs, experience in teaching. P working in MLR Institute of Technology, H a Assistant Professor in Department of Computer Science and Engineering. My specialized area is Knowledge and data engineering. I published 6 papers in this area, 6 are International Journals and two International Conference).

Dr. P. Bhaskara Reddy, B.E.(ECE), M.Tech., Ph.D., F.I.S.E.E., MCSI, MISTE, the Director MLR Institute of Technology is a young and dynamic professor of ECE, has 26 years of Teaching, Research and Administrative experience in Reputed Engineering Colleges and Industry. Recipient of Bharath Jyothi award in 2003 and Rastraprathiba award in 2004, Knowledge Award from Alumni of SVHCE for the year 2001, Published 1 Book (International Edition) "Information Technology in Technical Education – Economic Development by "LAMBERT Academic Publishing", Published 9 Laboratory Manuals, 65 Research papers at National and International Level on Education, Electronics Communication, I.T, Computer Networks, E-Commerce etc. Guided 5 Research Scholars for their Doctorates, about 50 M.Tech., M.C.A. and B.Tech projects and Conducted 10 National Level Technical Symposiums on various topics in Electronics & Communications, Computers etc.

