An Elegant Context Perceptive Page Rank Algorithm for Information Retrieval

Varun Singh[1] Tarun Dhar Diwan[2] Rohit Miri[3]

[1,2,3]Dr. CV Raman University, Bilaspur C.G.

Abstract— The World Wide Web contains the large amount of information sources and these are increasing tremendously. When the user searching the web for information retrieval, user may fetch irrelevant and redundant data causing a waste in user time and accessing time of the search engine. So narrowing down this problem by taking user interest and need from their behavior into account, have become increasingly important. Web structure mining plays an effective role in this approach. Some page ranking algorithms PageRank , Weighted PageRank are commonly used in web structure mining. The original PageRank algorithm search-query results independent of any particular search query. To yield more accurate search results for a particular topic, we have proposed a new algorithm, Elegant Context Perceptive Page Rank Algorithm on web structure mining that will show how the relevant pages of a given topic is better determined, as compared to the existing PageRank, Topic sensitive PageRank and Weighted PageRank algorithms.

Keywords - Page Rank, Information Retrieval, Weighted page rank.

I. INTRODUCTION

The World Wide Web [12] is the collection of information resources on the Internet that are using the Hypertext Transfer Protocol. It is a repository of many interlinked hypertext documents, accessed via the Internet. Web may contain text, images, video and other multimedia data. In order to analyze such data, some techniques called web mining techniques are used by various web applications and tools. Web mining describes the use of data mining techniques to automatically discover Web documents and services, to extract information from the Web resources and to uncover general patterns on the Web. Over the years, Web mining [14] research has been extended to cover the use of data mining and similar techniques to discover resources, patterns, and knowledge from the Web-related data (such as Web usage data or Web server logs).It is used to understand customer behavior, evaluate the effectiveness of a particular Web and help quantify the success of a marketing campaign. It is a rapidly growing research area. When a user makes a query from search engine, it generally returns a large number of pages in response to user queries. This result-list contains many relevant and irrelevant pages according to user's query. As user impose more number of relevant pages in the search result-list. To assist the users to navigate in the result list, various ranking methods are applied on the search results. The search engine uses these ranking methods to sort the results to be displayed to the user. In that way user can find the most important and useful result first.

Various link-based ranking strategies have been developed recently for improving Web-search query results. The HITS algorithm proposed in [22] relies on query-time processing to deduce the hubs and authorities that exist in a sub graph of the Web consisting of both the results to a query and the local neighborhood of these results. [4] augments the HITS algorithm with content analysis to improve precision for the task of retrieving documents related to a query topic (as opposed to retrieving documents that exactly satisfy the user's information need). [8] makes use of HITS for automatically compiling resource lists for general topics. The PageRank algorithm discussed in [23] pre-computes a rank vector that provides a-priori \importance" estimates for all of the pages on the Web. This vector is computed once, oine, and is independent of the search query. At query time, these importance scores are used in conjunction with query-specific IR scores to rank the query results. PageRank has a clear efficiency advantage over the HITS algorithm, as the query-time cost of incorporating the pre- computed PageRank importance score for a page is low. Furthermore, as PageRank is generated using the entire Web graph, rather than a small subset, it is less susceptible to localized link spam. In this paper, we propose an approach that (as with HITS) allows the query to influence the link-based score, yet (as with PageRank) requires minimal query-time processing. In our model, we compute offline a set of PageRank vectors, each biased with a different topic, to create for each page a set of importance scores with respect to particular topics. The idea of biasing the PageRank computation was suggested in [6] for the purpose of personalization, but was never fully explored. This biasing process involves introducing artificial links into the Web graph during the offline rank computation

The rest of this paper is organized as follows: a brief summary of related work is given in Section 2. Section 3 describes the proposed algorithm in detail. Finally we conclude in section 4.

II. RELATED WORK

With the increasing number of Web pages and users on the Web, the number of queries submitted to the search engines are also increasing rapidly. Therefore, the search engines needs to be more efficient in its process. Web mining techniques are employed by the search engines to extract relevant documents from the web database and provide the necessary information to the users. The search engines become very successful and popular if they use efficient Ranking mechanism. Google search engine is very successful because of its PageRank algorithm. Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some ranking algorithms depend only on the link structure of the documents i.e. their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining), while some use a combination of both i.e. they use content of the document as well as the link structure to assign a rank value for a given document. If the search

results are not displayed according to the user interest then the search engine will loose its popularity. So the ranking algorithms become very important. Some of the popular page ranking algorithms are discussed in the following section.

A. Citation Analysis

Link analysis is similar to social networks and citation analysis. The citation analysis was developed in information science as a tool to identify core sets of articles, authors, or journals of a particular field of study. "Impact factor" [10] developed by Eugene Garfield is used to measure the importance of a publication. This metric takes into account the number of citations received by a publication. The impact factor is proportional to the total number of citations a publications has. This treats all the references equally. Some important references which are referred many times should be given an additional weight. Pinski et al [11] proposed a model to overcome this problem called "influence weights" where the weight of each publication is equal to the sum of its citations, scaled by the importance of these citations.

The same principle is applied to the Web for ranking the web pages where the notion of citations corresponds to the links pointing to a Web page. This simplest ranking of a Web page could be done by summing up the number of links pointing to it. This favors only the most popular Web sites, such as universally known portals, news pages, news broadcasters etc. In the Web, the quality of the page and the content's diversity should also be considered. In the scientific literature, co-citations are between the same networks of knowledge. On the other hand, Web contains lot of information, serving for different purposes.

B. PageRank

The idea behind PageRank is that good pages reference good pages. Hence, pages that are referenced by good pages have higher PageRank. Although there are several formulations of PageRank, we use the random surf metaphor. Suppose that you are a user surfing the Web in a random fashion, such that, if you are in a page, with certain probability you get bored and leave the page, or you choose uniformly at random to follow one of the links on the page where you are (removing self links). Hence, the probability of being in page p is

$$PR(p) = \frac{q}{T} + (1-q)\sum_{i} \frac{PR(r_i)}{L(r_i)}$$

Where, T is the total number of pages, q is the probability of leaving page p (in the original work q = 0:15 is suggested), r_i are the pages that point to page p, and $L(r_i)$ is the number of links in page ri. These values can then be used as page ranking, and can be computed by an iterative algorithm converging quite fast, as we are interested in the ranking order rather than the actual ranking values. The term q is called damping factor as decreases exponentially link spamming based in sequences of links that return to a page.

C. The Weighted PageRank Algorithm

Weighted PageRank(WPR) algorithm which is an extension of the PageRank algorithm (Wenpu Xing and Ali Ghorbani, 2004) [1]. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance.

The importance is assigned in terms of weight values to the incoming and outgoing links are denoted as $W^{in}_{(m,n)}$and $W^{out}_{(m,n)}$ respectively $W^{in}_{(m,n)}$ as shown in equation is the weight of link(m, n) calculated based on the number of incoming links of page n and the number of incoming links of all reference pages of page m.

$$W^{in}_{(m,n)} = \frac{In}{\sum_{p \in R(m)} Ip}$$

$$W^{in}_{(m,n)} = \frac{On}{\sum_{p \in R(m)} Op}$$

Where and Ip are the number of incoming links of page n and page p respectively. R(m) denotes the reference page list of page m. $W^{out}_{(m,n)}$ is as shown in equation is the weight of link(m, n) calculated based on the number of outgoing links of page n and the number of outgoing links of all reference pages of m. Where On and Op are the number of outgoing links of page n and p respectively. The formula as proposed by Wenpu et al for the WPR is as shown in which is a modification of the PageRank formula.

$$WPR(n) = (1-d) + d\sum_{m \in B(n)} WPR(m) W^{in}_{(m,n)} W^{out}_{(m,n)}$$

D. The Topic Sensitive PageRank Algorithm

In Topic Sensitive PageRank [2], several scores are computed: multiple importance scores for each page under several topics that form a composite PageRank score for those pages matching the query. During the offline crawling process, 16 topic-sensitive PageRank vectors are generated, using as a guideline the top-level category from Open Directory Project (ODP). At query time, the similarity of the query is compared to each of these vectors or topics; and subsequently, instead of using a single global ranking vector, the linear combination of the topic-sensitive vectors is weighed using the similarity of the query to the topics. This method yields a very accurate set of results relevant to the context of the particular query.

With Topic Sensitive PageRank a set of ranking vectors are computed, as opposed to the single PageRank vector generated using standard PageRank. These vectors are biased using a set of representative topics, to capture the notion of importance with respect to a topic, indirectly specified through a user query and if available through user context also.

E. The Hyperlink Induced Topic Search Algorithm (HITS)

Kleinberg, 1999 [22] suggests a different method for ranking web pages through the introduction of hubs and authorities. The HITS technique is based on the premise that a sufficiently broad topic contains communities of pages. In order to rank pages each relevant page is assigned a hub and authority score. An authoritive page is a highly referenced topic-relevant page and a hub page is a page pointing to many authoritive pages. HITS starts by assembling a collection of Web pages C, which should contain communities of hubs and authorities pertaining to a given

topic t. It then analyzes the link structure induces by that collection, in order to identify the t-authoritative pages.

III. PROPOSED METHODOLOGY

PageRank and Weighted PageRank algorithms are used by many search engines but the users may not get the required relevant documents easily on the top few pages. With a view to resolve the problems found in both algorithms, a new algorithm called context perceptive page rank has been proposed which employs Web structure mining as well as Web content mining techniques. This algorithm is aimed at improving the order of the pages in the result list so that the user may get the relevant and important pages easily in the list.

Context Perceptive Page Rank is a proposed page ranking algorithm which is used to give a sorted order to the web pages returned by a search engine in response to a user query. This algorithm employs web structure mining as well as web content mining techniques. Web structure mining is used to calculate the importance of the page and web content mining is used to find how much relevant a page is? Importance here means the popularity of the page i.e. how many pages are pointing to or are referred by this particular page. Relevancy means matching of the page with the fired query. If a page is maximally matched to the query, that becomes more relevant. Context Perceptive Page Rank based on web structure mining that will show how the relevant pages of a given topic is better determined, as compared to the existing PageRank algorithm.

A. Algorithm: ECP PageRank

Input: Page P, Inlink and Outlink Weights of all backlinks of P, Query Q,
Output: Rank score
Step 1: Perceptive calculation:

```
// Input:    Text with n characters and Pattern with m characters
//Output: Index of the first substring of T matching P
    Compute function last
    i ← m-1
    j ← m-1
    Repeat
      If P[j] = T[i] then
        if j=0 then
          return i      // we have a match
        else
          i ← i -1
          j ← j -1
      else
        i ← i + m - Min(j, 1 + last[T[i])
        j ← m -1
    until i > n -1
    Return "no match"
```

Step 2: Rank calculation:

a) Find all backlinks of P (say set B).
b)

$$PR(P) = (1-d) + d[\sum_{V \in B} PR(V)Win(P,V)Wout(P,V)](CW + PW)$$

c) Output PR(P) i.e. the Rank score

IV. CONCLUSIONS

The rapid proliferation of World Wide Web has led the web content to increase tremendously. Hence, there is a great requirement to have algorithms that could list relevant web pages accurately and efficiently on the top of few pages. Mostly search engines used PageRank, Weighted Page Rank but users may not get required documents easily. With a view to resolve the existing problems, a new algorithm called Elegant Context Perceptive Page rank has been proposed which employs Web Structure mining. . For each Web page, compute an importance score per topic. At query time, these importance scores are combined based on the topics of the query and associated context to form a composite Page Rank score for those pages matching the query. This score can be used in conjunction with other scoring schemes to produce a final rank for the result pages with respect to the query. This algorithm will improve the order of web pages in the result list so that user may get the relevant pages easily. In future we endeavor to design an architecture which can continuously update and refresh the web information and update the repository periodically.

REFERENCES

[1] W. Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proc. ofthe Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.
[2] Taher H. Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No4, July/August 2003, 784-796.
[3] 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.
[4] N. Duhan, A. K. Sharma and K. K. Bhatia, "Page Ranking Algorithms:A Survey, Proceedings of the IEEE International Conference on Advance Computing, 2009.
[5] 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.
[6] M. Bianchini, M.. Gori and F. Scarselli, "Inside PageRank". ACM Transactions on Internet Technology, Vol. 5, Issue 1, 2005
[7] J. Cho and S. Roy, "Impact of Search Engines on Page Popularity". Proc. of the 13th International Conference on WWW, pp. 20-29, 2004.
[8] J. Cho, S. Roy and R. E. Adams, "Page Quality: In search of an unbiased web ranking". Proc. of ACM International Conference on Management of Data". Pp. 551-562, 2005.
[9] A. M. Zareh Bidoki and N. Yazdani, "DistanceRank: An intelligent ranking algorithm for web pages" Information Processing and Management, Vol 44, No. 2, pp. 877-892, 2008.
[10] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins,

"Mining the Link Structure of the World Wide Web", IEEE Computer Society Press, Vol 32, Issue 8 pp. 60 – 67, 1999.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing order to the Web". Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[12] J. Hou and Y. Zhang, "Effectively Finding Relevant Web Pages from Linkage Information", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, 2003.

[13] J. Dean and M. Henzinger, "Finding Related Pages in the World Wide Web", Proc. Eight Int'l World Wide Web Conf., pp. 389-401, 1999.

[14] R. Cooley, B. Mobasher and J. Srivastava, "Web Minig: Information and Pattern Discovery on the World Wide Web". Proceedings of the 9thIEEE International Conference on Tools with Artificial Intelligence, pp. (ICTAI'97), 1997.

[15] Sung Jin Kim and Sang Ho Lee, "An Improved Computation of the PageRank Algorithm", In proceedings of the European Conference on Information Retrieval (ECIR), 2002.

[16] Ricardo Baeza-Yates and Emilio Davis ,"Web page ranking using link attributes" , In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, PP.328-329, 2004.

[17] H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008.

[18] Fabrizio Lamberti, Andrea Sanna and Claudio Demartini , "A Relation-Based Page Rank Algorithm for. Semantic Web Search Engines", In IEEE Transaction of KDE, Vol. 21, No. 1, Jan 2009.

[19] Lian-Wang Lee, Jung-Yi Jiang, ChunDer Wu, Shie-Jue Lee, "A Query-Dependent Ranking Approach for Search Engines", Second International Workshop on Computer Science and Engineering, Vol. 1, PP. 259-263, 2009.

[20] NL Bhamidipati et al., "Comparing Scores Intended for Ranking", In IEEE Transactions on Knowledge and Data Engineering, 2009.

[21] Su Cheng,Pan YunTao,Yuan JunPeng,Guo Hong,Yu ZhengLu and Hu ZhiYu "PageRank, "HITS and Impact Factor for Journal Ranking", In proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering – Vol. 06, PP. 285-290, 2009.

[22] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[23] Larry Page. PageRank: Bringing order to the web. Stanford Digital Libraries Working Paper, 1997.

[24] Sergey Brin, Rajeev Motwani, Larry Page, and Terry Winograd. What can you do with a web in your pocket. In Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 1998.

