PAGE RANKING: SOME RECENT TRENDS AND FUTURE ASPECTS

Amit Sharma*
S.K.Dubey**
Abhishek Goyal***

ABSTRACT

Page rank is an algorithm which provides the importance of global estimate of web pages. Since, current search engine cannot rank the page of an individual user need and performance. In this paper we are presenting some algorithms on page ranking and their applications and also presenting limitations and their future development.

* CSE/IT Department, Venkateshwara Institute of Technology, Meerut

**CSE/IT Department, Amity University, Noida

***CSE/IT Department, Venkateshwara Institute of Technology, Meerut

INTRODUCTION:

The web is a highly distributed and homogeneous environment. A large number of web documents present various challenges for search engines. Recent search engine.s rank pages by combining traditional information retrieval techniques based on page content, such as word vector space with link analysis techniques based on the hypertext structure of the web, such as page rank and HITS. The page rank algorithms provide a global ranking of web pages based on their importance estimated from hyperlinks. For example, a link from page .A. to page .B. is Considered as if page .A. is voting for the importance of page .B.. So the number of links to page .B. increases, its importance increases as well. In Page Rank, not only the number of page in-links but their sources decide the importance of a page. In this scenario the global ranking of pages is based on the Web graph structure. Search engine such as Google utilize the link structure of the web to calculate the Page Rank value of pages. These values are then used to rank search results to improve precision. However, unlike flat document collection, the World Wide Web is hypertext and provides considerable auxiliary information on top of the text of the web pages, such as link structure and link text. The page rank algorithms attempts to provide an objective global estimate of web page importance However the importance of web pages is subjective for different users and thus can be better determined if the page rank algorithms takes into consideration user preference. The importance of a page depends on the different interest and knowledge of different people; a global ranking of a web page might not necessarily capture the importance of that page for a given individual user. Here we explore how to personalize page rank based on features readily available from page from URLs. For instance a user might favor pages from a specific geographic region, as may be revealed by internet domains. Likewise, topical features of the internet domains might also reflect user preference. A user might prefer pages that are more likely to be monitored by experts for accuracy and quality such as paper

published by academic institutions. Current search engines cannot rank pages based on individual user needs and preference.

SURVEY OF LITERATURE:

The idea of page rank was first introduced by Page, L. et al (1989) and has been studied by various researchers as a query- dependent ranking mechanism. There has been a great deal of work on academic citation analysis. Goffman (1971) has published an interesting theory of how information flow in a scientific community is an epidemic process. There has been a fair amount of recent activity on how to exploit the link structure of large hypertext system such as web. Pitkow completed his Ph.D. on characterizing World Wide Web ecology with a wide variety of link based analysis. In the order to address the above limitation of global page rank, we introduce a methodology to personalize page rank scores based on URL features such as internet domains. In the scenario, users specify interest profiles as binary features vectors where a features corresponds to a DNS tree node or node set. we pre compute Page Rank scores for each profile vector by assigning a weight to each URL weights , and used a t query time to rank result. We present promising preliminary results for a small experiment in which users where allowed to selected among time URL features combining the top two levels of the DNS tree leading to pre-computed page rank vectors. The idea of a personalized page Rank was first introduced by Page et al (1998) and studied by many research scholars as a query-depended ranking mechanism. If Personal preferences are based on n binary features; there are 2n different personalized page rank vectors for all user preferences .this requires an enormous amount of computation and storage facilities. In an attempt user to solve this problem, a method was introduced that computes only a limited amount of page rank vectors offline by Jeh and Widom (2003) . This method provides for a methodology where personalized page Rank vectors can be compute at query time for others possible user preferences. The main concern of the presented here to introduce a methodology for personalizing page Rank vectors based on URL features. To this end, we limit these choices of user preferences to topical and geographic features of internet domains. Techniques for

efficient and scalable calculation of page Rank scores are an area of very active research, while this important and relevant to the issue of personalized page Rank discussed. For the experiments presented here we use a collection from a relatively small crawl and it is not necessary to recomputed page Rank frequently. Therefore scalability is not discussed further in recent literature. Google has recently started beta-testing a personalized web search service based on a topical user profiles .It appears that user profiles are based on hierarchical topic directories (a.la open directory project, for details www.dmoz.org), however due to lack of *documentation* we are unable to discuss the similarities or differences between this work and the methodology proposed here. The definition of page rank has another intuitive basis in random walk on graphs the simplified version corresponds to the standing probability distribution of a random walk on the graphs of the web. This can be thought of as modeling the behavior of a random surfer. The random surfer simply keeps clicking on successive links at random. However, if a real web surfer ever gets into a small loop of web pages, it is unlikely that the surfer will continue in the loop forever. Instead, the surfer will jump to some other page. Page et al (2002) submitted their technical report citing page rank algorithm, Agogino and Ghosh (2002) gave the idea about increasing the page rank. Robert and Riggs (2001) developed an algorithm for automatically ranking. Clausen (2004) discussed about the cost of page ranking. Page and Brin (1998) give the idea about the anatomy of large scale hypertext web search engine. Kleinberg (1999) searched some authoritative sources in a hyperlinked environment. Deskin and srivastava (2003) presented temporal behavior of web usages. Acharya and Ghosh (2004) gave estimation of page ranking for missing data. Kamvar et al.(2003)defined adaptive method for computation of page rank. While Haveliwala (2002) discussed about topic sensitive page ranking.

SOME RECENT ALGORITHMS ON PAGE RANKS

(1) The Page Rank algorithm described by Lawrence Page and Sergey Brin in the several publications and it can be given as

$Pr(A)= (1-d) + d[Pr(T_1)/C(T_1)+...+Pr(T_n)/C(T_n)]$

1. Where Pr(A) is the Page Rank of page A,

2. Pr(Ti) is the page rank of pages Ti which links to page A

3. C(Ti) is the number of outbound links on page Ti and

4. d is the damping factor

From the results from Lawrence Page and Sergey Brin, the Page Rank does not rank websites as a whole but is determined for each page individually, Further the Page Rank of page A is defined by the page ranks of those pages which link to page A. The page rank of page Ti which links to page A does not influence the page rank of page A uniformly. Within the page rank algorithm, the page rank of page T is always weighted by the number of outbound links C(T)on page T. This implies more outbound links a page T, the less will a page A benefit from a link to it on page T. The weighted Page Rank of page Ti is then added up. The outcome of this page is that an Additional inbound link for page A will always increase page A.s Page Rank. Finally, the sum of the weighted Page Rank of pages Ti is multiplied with a damping factor d which can be set between 0 and 1. There by, the extend of Page Rank benefit for a page by another page linking to it is reduced.

(2) In the second version of the algorithm, the page rank of A is given by,

$$Pr(A)=(1-d)/N + d[Pr(T1)/C(T1)+...+Pr(Tn)/C(Tn)]$$

Where N is the total number of all pages on the web, The second version of the algorithm does not differ fundamentally from the first one. Regarding the random surfer model, the second version's page rank of a page is actual probability for a surfer reaching that page after clicking on many links. The page ranks then from a probability distribution over web pages, so the sum of all pages. Page ranks will be one. Contrary, in the first version of the algorithm the probability for the random surfer reaching the page is weighted by the total number of web pages. So, in this version page rank is an expected value for the random surfer visiting a page, when he restarts this Procedure as often as the web has pages. If the web has 100 pages and a page had a page rank value 2, the random surfer would reach the page in an average twice if he restarts 100 times.

As mentioned above, the two versions of the algorithm do not differ fundamentally from each other. A page which has been calculated by using the second version of the algorithm has to be

multiplied by the total number of web pages to get the according page rank that would have been calculated by using the first version

Personalized Page Rank Vectors

Personalized Page Rank vectors provide a ranking mechanism which in true creates a personalized view of the web for individual users. The computation of personalized page rank vectors is done prior to search time. When calculating the page Rank vectors, predefined user profiles are taken into consideration. We use the following recursive definition for personalized page rank computation.

$$R_u(P) = (1-d) + d\sum_{Q:Q\to P} \frac{Wu(q)Ru(q)}{S:Q\to S}$$

Where U is the user profile, d is the traditional jump property (for damping factor), the sum over page q that links to p has each element normalized by the number of out links from page q based on profile U.

FUTURE WORK:

One must go with some algorithms for more accurate page display as per the requirement of user's need. Since the search engine giving the data which have a large amount of unwanted or none required.

REFERENCES:

1. Acharya,S. and Ghosh, J.(2004): **Outline Estimation for Page Rank computation under Missing Data**. Proceedings, 13th International world wide web conference. pp486-487.

2. Clausen, A. (2004): **The cost of page rank**. In Procedings of International conference on agents, web technologies and Internet commerce, Gold Cost Australia

3. Eiron,N., McCurley, K. and Tomlin, J.(2004):**Ranking the Web frontiers**. 13th International world wide web ACM Press. Pp309-318.

4. Stuart E. Middleton et al.(2004): **Ontological user profiling in recommender systems**. ACM Transactions on Information systems, Vol 22(1)pp1-4

5. Deskin, P. and Srivastava, J.(2003): .**Temporal wehavior of web usage**., AHPCRC technical report, August,

6. Jeh, G. and Widom, J.(2003): **Scaling personalize Web search**. In: Proceedings 12th International world wide web conference.

7. Kamvar, S.D, Haveliwala,T.H., Manning,C.D. and Golub, G.H.(2003): **Extrapolation methods for accelerating the computation of Page Rank**. In: Proceedings 12th International world wide web conference.

8. Kamvar, S.D, Haveliwala,T.H. and Golub, G.H.(2003): **Adaptive Methods for the Computation of Page Rank**. Technical Report, Stanford University.

9. Kamvar, S.D, Haveliwala,T.H., Manning,C.D. and Golub, G.H.(2003): **Exploiting the block structure of the web for computing Page Rank**. Technical Report, Stanford University

10. Agogino, A. and Ghosh,J. (2002): **Increasing page rank through reinforcement learning**. In Proceedings of Intellegent Engineering Systems Through Artificial Neural Networks.vol-12, ASME Press, pp27-32

11. Haveliwala,T.(2002): **Topic-sensitive Page Rank**. 11th International World Wide Web ACM Press.

12. Page L., Brin, S. Raghavan, P. and Upfal, E.(2002): **The Page Rank Citation Algorithm Bringing order the Web**. Technical report, Stanford Digital Library Technologies project.

13. Robert Wilensky and Tracy Riggs (2001): **An algorithm for automatically rating reviewers**. In: Proceedings of the First Joint Conference on Digital Library.

14. Haveliwala,T.(1999): **Efficient computation of Page Rank**. Technical report, Stanford Database group.

15. Kleinberg, J.(1999): **Authoritative sources in a hyperlinked environment**. Journal of the ACM 46,pp604-632

16. Brin, S. and Page, L.(1998): **The anytomy of a large-scale hypertextual Web search engine**. Computer network 30,pp107-117

17. Page, L., Brin S., Motwani, R. and Winogerad, T.(1998**): The Page Rank citation ranking: Bringing order to the Web**. Technical report, Standford University Database Group.

18. Motwani, R. and Raghavan, P.(1995): **Randomized Algorithms**. Cambridge University, Press.

19. Salton, G. and McGill, M.(1983): **An introduction to modern information Retrieval**. McGraw-Hill, New York

20. Goffman,W.(1971): **A mathematical method for analyzing the growth of a scientific discipline**. Journal of ACM, vol-18(2),pp173-185.

21. Van Rijsbergen**, C**.(1970).: **Information Retrieval**. Butterworths, London Second edition.