International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles are freely available online:http://www.ijoer.in
Vol.1.Issue.1:2013

RESEARCH ARTICLE

SURVEY ON PAGERANK ALGORITHMS USING WEB-LINK STRUCTURE

SOWMYA.M[1], V.S.SREELAXMI[2], MUNESHWARA M.S[3], ANIL G.N[4]
Department of CSE, BMS Institute of Technology, Avalahalli, Yelahanka, Bengaluru-560064
[1]sowmyam.upadhyaya@gmail.com, [2]vss57@yahoo.com, [3]muniaitckm@gmail.com, [4]anilgrama@yahoo.co.in

Article Received: 03/05/2013 | Revised on: 04/05/2013 | Accepted on: 24/05/2013

MUNESHWARA M.S

ABSTRACT
Pagerank vector for ranking the search-query results, which made use of link structure of the Web, to get the importance of Web pages, particularly independent of any search query. To get correct search results, we propose calculating a set of PageRank vectors, biased with a set of archetypical topics, to capture more correctly the notion of prominence with respect to a particular topic. By taking these biased PageRank vectors we compute query-specific rank scores for web pages at query time, it is shown that we can compute more accurate importance score than with a single PageRank vector. We compute the topic-sensitive PageRank scores for pages for normal keyword search queries, sufficing the query using the topic of the keywords. PageRank scores using context in which the query appeared. For better ordering of web pages compute an associated PageRank algorithm for search engines to get quality results by scoring based on relevance between web documents. The modified PageRank algorithm creates certain ordering using relevance than the original one, and reduces the query time overhead of topic-sensitive PageRank

KEYWORDS-HITS, Hyperlinks, MFTS, ODP, Backlinks,PageRank.

INTRODUCTION

PageRank is an important factor to score web documents, but it is biased by link spam easily, so that search engines like Google have to evaluate alternative factors to adjust the result of ranking. In this paper, we will discuss the accuracy and efficiency of different page ranking algorithms, and propose an associated PageRank to retain the relevance between web pages and to prevent the bias of link spam.

Page Rank is something like an author-to-author voting system which accumulates the weighted inbound links returned from other websites. The weight depends on the number of outbound links and Page Rank value of the source web pages. The original Page Rank algorithm is described below [6].

$$PR(V_i)=(1-d)+d\times \sum_{j\in In(V_i)} \frac{PR(V_j)}{Out(V_j)}$$

d: damping factor 0~1, normally set to 0.85
PR(*iV*): Page rank of page *iV*
In(*iV*): the number of inlink of page *iV*
Out(*jV*): the number of outlink of page *jV*

If there are three pages {A, B, C}, and A links to B and C, and C links to A as illustrated in Figure 1. The PageRank of A will be 0.43444227, and PageRank of B and C will be 0.334638, if damping factor is 0.85 and initial PageRank is 1. Figure 2 shows the PageRank distribution with several iterations. The PageRank values will converge to one value after certain iterations with any initial value, and we will

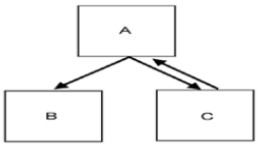

Figure 1. Page link relation

Most organizations request their commercial web pages to be ranked as high as possible in the results returned by search engines.

These requests can be achieved by link spam and keyword spam if the ranking algorithm cannot detect artificial tricks. With the increasing demand of web searches, the modifications of Page Rank have become more critical, and the precision of Page Rank can affect satisfaction of search engine results pages.

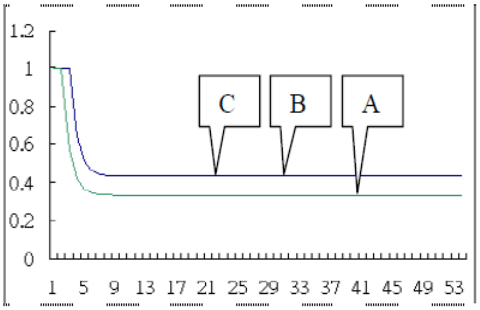

Figure 2. PageRank distribution

SIMPLE PAGERANK

The Search Engine has a list with all the websites it indexes. For the indexing it uses so called crawlers to 'read' through the websites. Now you have a huge database and a list of websites thatcontain the word you might search for. (Simplified!)

But in which order do you want the results to be displayed? All methods used so far were very vulnerable to attacks, and there is a strong urge to manipulate the position in which a website is displayed in the Search Engine result. Those methods used for instance the number of appearances of the word, or the word's context, its size relative to the surrounding text, etc. But you can easily see how you can corrupt these methods. Just insert words, which are often used in queries, like celebrity names, into the meta text and your site will be under the top sites for many queries (but it's possibly not a page the client searched for). In fact, the 'Page Rank Paper' mentions that in 1998 only one of the top four commercial Search Engines finds itself if you search for its name and returns the result in the top ten.

In 1998 Larry Page and Sergey Brin proposed a rather new approach which used the web's underlining link-structure to grade the relevance of each website

***A SIMPLE EXAMPLE*:** To begin with we don't know the Ranks of the web pages so we assign them one. For simplicity we will choose the number 1. So the diagram with Rank on it becomes.

The damping factor (d) basically says that a page cannot vote another page to be as equally important as it is.

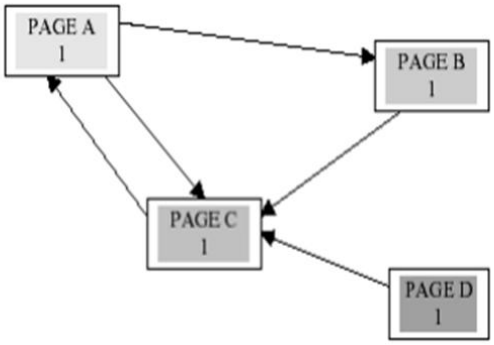

Figure 3. Page Rank example

- Page A first, the amount of Rank available to pass on, after dampening it down, is 1*0.85 = 0.85. There's two links out, so at the end of the process we're going toadd 0.425 to Page B's Rank and 0.425 to Page C's Rank.
- On to Page B. It has just one link. So it'll pass on 1*0.85 = 0.85 to Page C.
- Page C also only has one link. So it'll pass on 1*0.85 = 0.85 to Page A.
- Page D has one link so it passes 0.85 to Page C.
- Now we can add all those totals on to all the pages. The new Rank totals show how important Page C is.

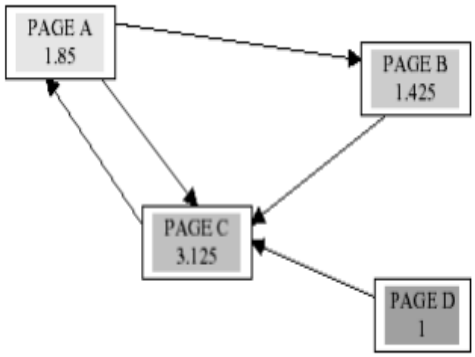

Figure 4.Pagerank example 1st iteration

TOPIC-SENSITIVE PAGERANK

Rungsawang and et al. introduce a PageRank computation to un-bias the link farm effect. It is a good algorithm if the link farm can be identified efficiently, but it is a more complicate situation in the real world. Havellwala proposes a topic-sensitive PageRank algorithm to evaluate web pages

with consideration of category relevance [10]. This modification can approach more precise scores of web pages, but the computation complexity will be a heavy load to index world-wide documents and reduce the efficiency in the query time. Al-Saffar and et al. follow the Havellwala's idea and claim a new approach for personalization without relying on the web link structures.

Topic-sensitive PageRank uses ODP-biasing (Open Directory Project) and query-time importance scores to evaluate pages importance [7].

ASSOCIATED PAGERANK

Associated PageRank algorithm is almost the same as original PageRank in spirit. The difference is that associated PageRank calculate the page relevance of outbound links using the most frequent terms sets and score the weights of PageRank values according to the relevance.

A document space consists of document *Di* which is identified by a set of frequent terms *Tj*; where the terms may be weighted by *Wj* according to the importance in each document. In order to measure the relevance between different documents *Dp* and *Dq*, the *m* most frequent terms are retrieved from documents.

In the first scenario, the most frequent term sets (MFTS) from different documents can be compared with the most frequent term sets of web pages in ODP (Open Directory Project) categories; therefore, the relevant degree of target documents for certain category can be calculated. In the other scenario, the content relevance can be determined directly by the semantic distance value of the most frequent term sets from the document *Dp*and *Dq*. Associated PageRank is implemented using two algorithms:

- Algorithm with category MFTS
- Algorithm without category MFTS

HITS ALGORITHM

Hyperlink-Induced Topic Search (HITS) (also known as hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It was a precursor to PageRank. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs.

The scheme therefore assigns two scoresfor each page: its authority, which estimates the value of the content of the page and its hub value, which estimates the value of its links to other pages [2].

HITS algorithm is in the same spirit as Page Rank. They both make use of the link structure of the Web graph in order to decide the relevance of the pages. The difference is that unlike the Page Rank algorithm, HITS only operates on a small sub graph (the seed SQ) from the web graph. This sub graph is query dependent; whenever we search with a different query phrase, the seed changes as well. HITS ranks the seed nodes according to their authority and hub weights. The highest ranking pages are displayed to the user by the query engine.

EXPERIMENTS

We follow the example of section I to inspect the Page Rank distribution of different Page Rank algorithms. In this example described in Figure 1, all pages share the Page Rank values from outbound links, so the traditional Page Rank values of page B and C are equal to 0.334638. If the page B is relevant to page A, and page C is just a link exchanging page of page A, the PageRank values of page B and C should not be the same. According to the traditional Page Rank algorithm, all outbound pages get the equal Page Rank no matter what the contents are. After setting $RAB=8.00\times10^{5-}$, and $RAC=1.00\times10^{7-}$, we get the associated Page Rank distribution as Figure 3 with $PR(B)=3.86\times10^{1-}$, $PR(A)=2.77751\times10^{1-}$, and $PR(C)=1.5\times10^{1-}$. We find the values converge faster than the traditional PageRank, and page B get the fair value since it is more relevant to A than C.

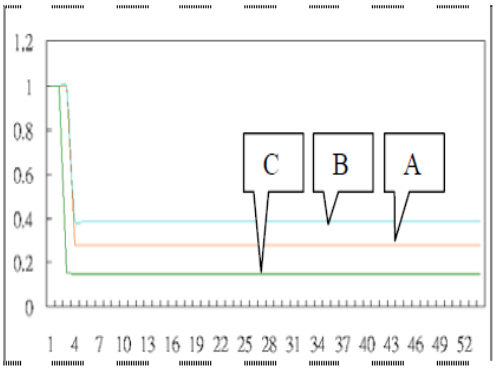

Figure 5: Associated PageRank distribution of page A, B, C

We setup two types of datasets: (1) random web pages from dmoz.org in different categories, and (2) web links of a certain page. From the first datasets, we calculate the relevance values of these random web pages comparing with http://www.cra.org to identify if the relevance values

of web pages in the same category of cra.org are higher than the others.

Figure 6 shows that only three values are greater than 8.00×106−, which are the relevance values of web pages in the same category with cra.org. The other relevance values in the different categories with cra.org are less than 8.00×106−. We conduct the same type of experiments for various websites and find that the pages are not relevant if the relevance values are far below 1.00×105−.From the second datasets, we calculate the relevance values of outbound links of cnn.com, and find that above 95% are greater than $1.00\times10^{5-}$. Figure 7 shows that 95% pages are relevant pages in cnn.com.

In order to investigate another characteristic of relevance values, we collect 50 personal blogs with link exchanges, and find that relevance values are relatively low for this type of websites.

For the traditional PageRank algorithm, all the outbound links share the PageRank value equally, but we can use the relevance value to adjust the weights according to the relevant degrees to get more precise PageRank. Moreover, by the analysis of the relevance values of outbound links, we can also detect if the link spam problem exists in web pages. The precision of the associated PageRank can also be improved by stop words selection and word stemming techniques.

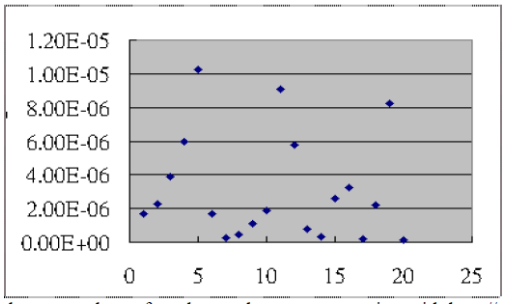

Figure 6. Relevance values of random web pages comparing with http://www.cra.org

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. At query time, these importance scores are combined based on the topics of the query to form a composite PageRank score for those pages matching the query. 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.

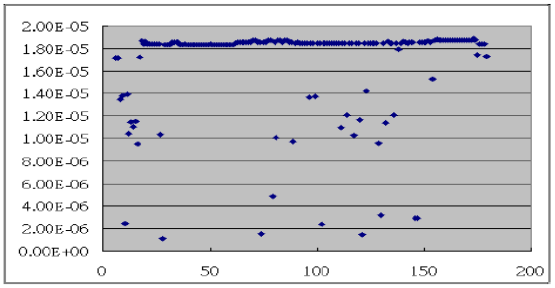

Figure 7. Relevance values of outbound linking pages of http://www.cnn.com

As the scoring functions of commercial search engines are not known, in our work we do not consider the effect of these other IR scores. We believe that the improvements to PageRank's precision will translate into improvements in overall search rankings, even after other IR-based scores are factored.

The most frequent term sets (MFTS) from different documents can be compared with the most frequent term sets of web pages in ODP (Open Directory Project) categories; therefore, the relevant degree of target documents for certain category can be calculated. In the other scenario, the content relevance can be determined directly by the semantic distance value of the most frequent term sets from the document *Dp* and *Dq*.

We setup two types of datasets: random web pages from dmoz.org in different categories, and web links of a certain page. From the first datasets, we calculate the relevance values of these random web pages comparing with http://www.ncsu.org to identify if the relevance values of web pages in the same category of cra.org are higher than the others.

We see that only three values are greater than 8×16−, which are the relevance values of web pages in the same category with cra.org. The other relevance values in the different categories with cra.org are less than 0.8×16−.

We conduct the same type of experiments for various websites and find that the pages are not relevant if the relevance values are far below 0.1×15−.

From the second datasets, we calculate the relevance values of outbound links of cnn.com, and find that above 95% are greater than $1.00\times10^{5-}$. It shows that 95% pages are relevant pages in cnn.com. Shows that all pages are relevant in bbc.co.uk and all values are above $12.00\times10^{5-}$. It shows the relevant values of outbound links of ieee.org are clustered around.7×16−. In order to investigate another characteristic of relevance values, we collect 50 personal blogs with link exchanges, and find that relevance values are relatively low for this type of websites. The relevance

values of outbound links of one certain personal blogs indicates there are many irrelevant outbound links in this personal blog.

CONCLUSION AND FUTURE WORK

PageRank is a global ranking of all web pages based on their locations in the web graph structure. PageRank uses information which is external to the web pages – backlinks. Backlinks from important pages are more significant than backlinks from average pages. The structure of the web graph is very useful for information retrieval tasks.

We can sightsee numerous ways of improving the approach for associated PageRank.. Another area of investigation is to optimize of the best set of basis topics. For instance it may be advisable to use a better-grained set of topics, perhaps using the second or third level of the Open Directory hierarchy, rather than simply the top level. We can alsooptimise the selection on Most Frequent Term Set (MFTS). However, the space required for the storage of data is much higher. Another aspect we can look into would be the memory optimization.

ACKNOWLEDGMENTS

The authors would like to thank the editor, mysterious reviewers for their valuable suggestions that appreciably improved the quality of this paper, especially to Dept. of CS&E for their constant encouragement and providing us the overwhelming support and guidance to write this paper. Finally also thankful our B M S Educational Trustees, DearPrincipal Dr. S Venkateswran ,Teaching Faculties, and Non- Teaching faculties of Department of CSEB M S Institute of Technology, Yelahanka,Bengaluru, Karnataka, India.

REFERENCES

[1] Kritikopoulos, M. Sideri and I. Varlamis, Wordrank: A Method for Ranking Web Pages Based on Content Similarity, BNCOD '07. 24th British National Conference on Database, pp. 92-100, 2007.

[2] C.H. Li and K.Q. Lv, Hyperlink Classification: A new approach to improve pagerank, 18th International Conference on Database and Expert Systems Applications, pp. 274-277, 2007.

[3] D. Cai, X. He, J. Wen, and W.Y. Ma, Block-level Link Analysis, 27th Annual International ACM SIGIR Conference, pp. 440-447, 2004.

[4] L. Page, S. Brin, R. Motwani and T. Winograd, The pageRank citation ranking: bringing order to the web, Technical Report, Stanford InfoLab, 1998.

[5] S. Brin and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[6] The Google Search Engine: Commercial search engine founded by the originators of PageRank. http://www.google.com/.

[7] The Open Directory Project: Web directory for over 2.5 millionURLs.http://www.dmoz.org/.
http://searchenginewatch.com/sereport/99/11-google.html

[8] A.L.Barabasi and R. Albert, Emergence of scaling in random networks, Science Magazine, Vol. 286. no. 5439, pp. 509-512, 1999.

[9]A.Rungsawang, K.Puntumapon, B. Manaskasemsak, Un-biasing the link farm effect in PageRank computation, 21th International Conference on Advanced Networking and Applications, pp. 924-931, 2007. Journal of Convergence Information Technology Volume 5, Number 8, October 2010

[10] T.H. Havellwala, Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, Issue 4, pp. 784-796, 2003.

[11] T.H. Havellwala, Topic-sensitive PageRank, International World Wide Web Conference, pp.517-526,2002