2010 Fourth International Conference on Network and System Security

An improved PageRank algorithm: immune to spam

Bing-Yuan Pu, Ting-Zhu Huang, Chun Wen
School of Mathematical Sciences
University of Electronic Science and Technology of China
Chengdu, PR China
Email: skypuby@163.com, tingzhuhuang@126.com, wchun17@163.com

***Abstract*—As Google claims on its webpage, PageRank™ is the heart of software and continues to provide the basis for all of web search tools. In this algorithm, one page's PageRank value is divided evenly among all its outlinks. This paper discusses the value not-even distributed question, and puts forward an improved PageRank algorithm. By illustrating examples, we verify the effectiveness of our new algorithm and especially immunity to electronic spam.**

***Keywords*-PageRank; Not-even-distribution; spam;**

I. INTRODUCTION

With the rapid growth of Internet, the WWW provides an important channel for user to obtain useful information. Whereas, there are more than eight billion web pages [1], so supplying information to user's content, is a challenging work. Fortunately, there are various web search engines, about 3500 different search engines [11], [12], such as Google, Yahoo, AltaVista, etc., by which we can find valuable information efficiently. Notess [14] and Ward [21] found that Google is unique in its focus on developing the perfect web search engine that understands accurately what users mean and gives them back precisely the desired information. Google's simple but effective algorithm, PageRank, ensures its leading. PageRank, a link structure-based algorithm formulated by Brin and Page [16], [4], gives a rank list of importance of pages related to user's query terms. This algorithm follows rules as [4]: a link from page A to page B is a vote from A to B; highly linked pages are more important than pages with few links. In this algorithm the rank score of a page p, named page's PageRank value (PR), is evenly divided among its outlinks. Because the huge size of actual WWW, an approximate iterative computation is usually applied to calculate the PR. Giving initial starting values, the PR distributed to the outlinks of page p are in turn used to calculate the PR of the pages, to which page p is pointing and the PR of all pages are then calculated in computation circles. Hence PR of page A is defined [4] as

$$PR(A) = \frac{1-d}{N} + d \times \sum_i \frac{PR(T_i)}{\|T_i\|}.$$

Here N is the total pages in WWW, $d \in (0,1)$ is a damping factor, and $\|T_i\|$ is the number of outlinks from T_i to A. In this algorithm, as each T_i outlinked to A contributes its PR to A, its PR is assigned averagely to all its outlinks (page A inside). This method can be carried out easily. But its weakness is obvious: in all the outlinks from A, some links may be more important than other pages, and so the more important pages should receive higher votes from A, not totally equivalent as the less important ones do. All pages outlinked from A get equal PR assignment from A, regardless of their importance issue, thus not only influences the ranking quality, but also provides chances for spam or speculating business activity as one company intends to enhance hyperlink to his home-page for advertisement. There are many improved algorithms about PageRank [17], [3], [10], [5], [13], [18], [19], [20]. They mainly focus on algorithm itself, including convergence, iteration, damping factor, hyperlink structure, etc. Discussion about PR average-distribution, mentioned above, appears lacking in this area. It is gratifying that [22] attempts to resolve this problem by introducing the number of inlinks and outlinks to weigh page's popularity. This method verifies it's more efficient than original PageRank algorithm [4]. But in our opinion, page's popularity shouldn't depend only on its number of links, obviously spam is an typical example of most convincing. In view of this, we improve the original PageRank algorithm [4], and put forward a new algorithm. When computing PR, contributed to A from T_i, we not only base on the outlinks number of page T_i, but the more emphasis on the PR of its outlink pages from T_i. Further, page T_i does not divide its PR among its outlinks mean, viceversa, does it according to their weights. In the end, we demonstrate an interesting phenomenon: our new PageRank algorithm can filter out spam, but the original PageRank algorithm is blinded. As known to all, identifying and preventing spam has been recognized as one of the top challenges in the search engine industry[8]. A lot of recent work mainly focuses on content or link analysis in their ranking shemes[2], [6], [7], [15]. For our work in this context, it carries it point of filtering out spam basically by link analysis, including number of links and value of linked object, which provides heuristic reference for consequent research.

The rest of this paper is organized as follows. Section 2 presents the new algorithm. Section 3 illustrates our algorithm by examples. Conclusively remarks are finally drawn in section 4.

978-0-7695-4159-4/10 $26.00 © 2010 IEEE
DOI 10.1109/NSS.2010.12

II. The improved PageRank algorithm

In the following discussion, let a target web page be A, we assume there are m pages $(B_1, B_2, ..., B_m)$ pointing to A. Moreover, we denote all pages outlinked from B_i are C_{ij} (its number is $||B_i||$). The original PageRank algorithm, as described by Page and Brin[4]is given by

$$PR_O(A) = \frac{1-d}{N} + d \times \sum_{i=1}^{m} \frac{PR(B_i)}{\|B_i\|}. \quad (1)$$

In regard to PageRank contribution from B_i to A, namely $\frac{PR(B_i)}{\|B_i\|}$, this algorithm obeys the equal distribution law among it's outlinks from B_i. In our new algorithm, we follow the principle: the highly linked pages are more important than pages with fewer links and we especially emphasize the law: when outlinked from more important page, more PageRank assigned to it, meanwhile, the less important pages get less page rank share.

By iProspect [9], an famous Search Engine Marketing company, "Key among the findings relating to the current search engine user community is that 62% of search engine users click on a search result within the first page of results, and a full 90% of search engine users click on a result within the first three page of search results." Since users mainly view the former pages when retrieve information in WWW, so we should raise the true important page's PR, and lower the "fake" important page's PR, and then establish the true important page in the front position in search pages and lag the less important pages behind, especially the spam included. Thus in our new algorithm, the PR of page A is calculated as follows:

$$PR_N(A) = \frac{1-d}{N} + d \times \sum_{i=1}^{m} \frac{PR_N(A)}{\frac{\|B_i\|}{\sum_{j=1} PR_N(C_{ij})}} \times PR_N(B_i). \quad (2)$$

There, $\frac{PR_N(A)}{\frac{\|B_i\|}{\sum_{j=1} PR_N(C_{ij})}}$, is the weight, page A in all the outlinks from B_i,

$$\frac{PR_N(A)}{\frac{\|B_{ij}\|}{\sum_{j=1} PR_N(C_{ij})}} \times PR_N(B_i),$$

is the part page B_i contributes to A. This algorithm intends to highlight the important pages, on the contrary play down the less important pages, especially the spam and links deliberately from company to propagandize itself. Obviously this idea complies with the PageRank$^{\mathrm{TM}}$ main rule: highly linked pages are more 'important' than pages with few links, and backlink from high PR-pages counts more than from low PR-pages [4].

Difference between original PR algorithm and the new PR algorithm are not-ignoring, and are discussed in section 3.

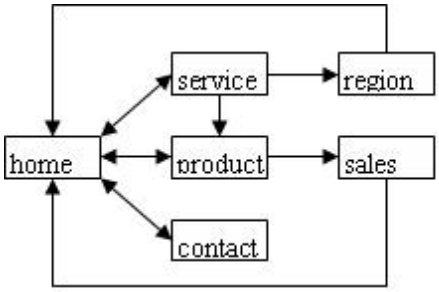

Figure 1: Hyperlink diagram with six pages.

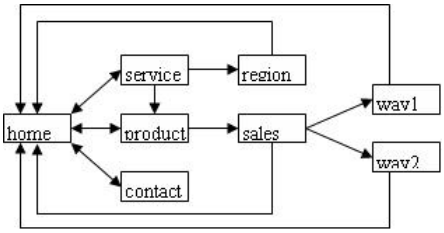

Figure 2: Hyperlink diagram with eight pages-the first.

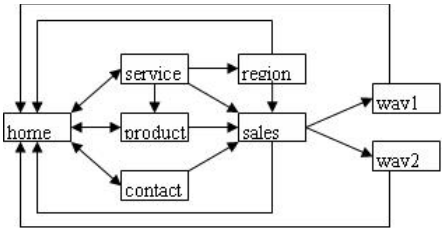

Figure 3: Hyperlink diagram with eight pages-the second.

III. Experiments

A. Experiment method

This experiment aims to confirm whether the two algorithms share the same page rank, and especially whether the new PageRank algorithm (NPR) has improved the original PageRank algorithm (OPR).

In this section, we present three small and simple examples to disclose the difference between the OPR and NPR. Figure 1 shows a hyperlink with only six pages. Figure 2 and Figure 3 show 8-page tiny web, with different link structure. As usual, we take d, the damping factor, the suggestion value 0.85 [4]. The iteration begins with each initial value equal to one. By 100 times iteration, the precision is within 10^{-8}. Table 1-3 list the computer result.

OPR			NPR		
Rank	Web	PageRank	Rank	Web	PageRank
1	Hom	0.391	1	Hom	0.441
2	Ser	0.136	2	Ser	0.111
3	Pro	0.174	3	Pro	0.262
4	Con	0.136	4	Con	0.111
5	Reg	0.063	5	Reg	0.029
6	Sal	0.099	6	Sal	0.046

Table I: *Diffrence in PageRank between the original PageRank algorithm and the new PageRank algorithm for Fig. 1.*

OPR			NPR		
Rank	Web	PageRank	Rank	Web	PageRank
1	Hom	0.369	1	Hom	0.441
3	Ser	0.123	3	Ser	0.100
2	Pro	0.158	2	Pro	0.263
3	Con	0.123	3	Con	0.100
6	Reg	0.054	6	Reg	0.021
5	Sal	0.086	5	Sal	0.035
7	Way1	0.043	7	Way1	0.020
7	Way2	0.043	7	Way2	0.020

Table II: *Diffrence in PageRank between the original PageRank algorithm and the new PageRank algorithm for Fig. 2.*

OPR			NPR		
Rank	Web	PageRank	Rank	Web	PageRank
1	Hom	0.318	1	Hom	0.422
4	Ser	0.109	3	Ser	0.100
3	Pro	0.132	2	Pro	0.242
4	Con	0.109	3	Con	0.100
8	Reg	0.042	8	Reg	0.021
2	Sal	0.162	5	Sal	0.072
6	Way1	0.065	6	Way1	0.022
6	Way2	0.065	6	Way2	0.022

Table III: *Diffrence in PageRank between the original PageRank algorithm and the new PageRank algorithm for Fig. 3.*

B. Experiment results and discussion

Table 1-3 indicate our original intention: upgrading the PageRank of more important pages and lower the ones of less important pages. Their results are schematized in Figure 4-6. The shapes in Figure 4-5 by the two algorithm are similar. Both algorithms have similar page rank. But when comparing Figure 5 and Figure 6, which imply similar link structure, we can easily find an exciting phenomenon. Many links are deliberately added to page 'sales' in Figure 3, by OPR algorithm, thus accordingly puts up the PageRank of page 'sales', arranging at second. In sharp contrast, by NPR algorithm, page 'sales' with more links although, but gets less PageRank value, arranging at fifth, than page 'service', 'product' and 'contact', because the latter get links from the most important page, 'home'. Imaging if the intended links to page 'sales' are caused by spam, by our new algorithm its PageRank does not get desired promotion, but still gets its 'true' PageRank. Thus, the nonsignificant pages, spam for example, have less page rank and have to bear disadvantage of being neglected by surfer, for its lower position in searching web. Nowthat, as known to all, Web spam pages use various techniques to achieve higher-than-deserved rankings in a search engine's results. These contrasts verify our new algorithm's feasibility and effectiveness, especially its capability to filtering out spam.

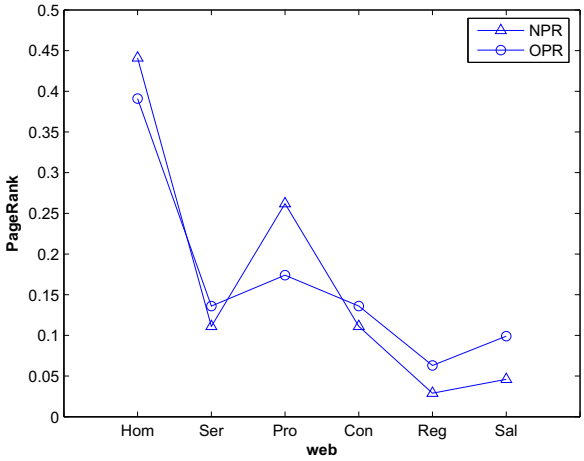

Figure 4: Difference in PageRank between the original PageRank algorithm and the new PageRank algorithm for Fig. 1.

IV. Conclusion and open problems

In this paper, we have discussed the not-equal-distribution problem: the outlinks from page A don't share averagely its PageRank. An new PageRank algorithm is pulled into to improve the original PageRank algorithm. By illustrating examples, we verify the new algorithm's effectiveness and its immunity to spam. There are several open problems, which should deserve some attentions. Since our algorithm is implemented at a partly cost in increased computation complexity, how about the convergency speed and system reserve of our new algorithm? And can we syncretize

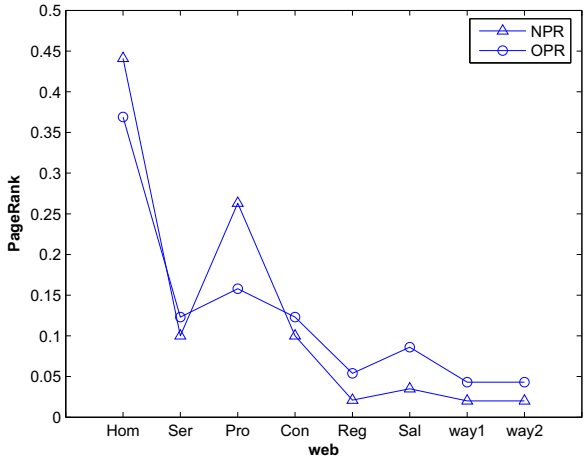

Figure 5: Difference in PageRank between the original PageRank algorithm and the new PageRank algorithm for Fig. 2.

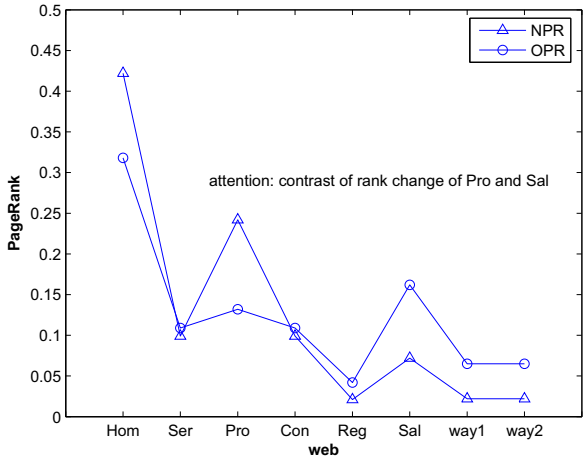

Figure 6: Difference in PageRank between the original PageRank algorithm and the new PageRank algorithm for Fig. 3.

our algorithm and other relevant acceleration methods in convergence? Meanwhile, we may think of combing link and content analysis in our algorithm to filter spam. These problems are our expectation of next work.

ACKNOWLEDGMENT

This research was supported by NSFC (10926190, 60973015) and Scientific Research Fund of Sichuan Provincial Education Department(2008zb068).

REFERENCES

[1] R. Baeza-Yates, F. Saunt-Jean, C. Castillo, Web structure, dynamics and page quality. *Proc 9th International Symposium on String Processing and Information Retrieval(SPIRE2002)*, p117-130.

[2] A. A. Benczur, K. Csalogany, and T. Sarlos, Link-based similarity search to fight web spam, *In Proc. AIRWeb*, 2006.

[3] C. Brezinski, et al., Extrapolation methods for PageRank computations, *C.R. Math. Acad. Sci. Paris*, 340(2005): 393-397.

[4] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine. *Proc 7th International World Wide Web Conference (WWW7)* , April(1998), p107-117

[5] H. H. Fu, K. J. L. Dennis, H. T. Tsai, Damping factor in Google page ranking, *Appl.Stochastic Models Bus. Ind.*, 22(2006): 431-444.

[6] Z. Gyongyi, H. Garcia-Molina, Web spam taxonomy, *In Workshop on Advers. Inf. Retrieval on the Web.*, 2005.

[7] Z. Gyongyi, H. Garcia-Molina, and, J. Pedersen, Combating web spam with trustrank, *In Proc. 30th VLDB.*, 2004.

[8] M. Henzinger, R. Motwani, and, C. Silverstein, Challenges in web search engines, *SIGIR FORUM.*, 36(2)(2002): 11-22.

[9] iProspect, iProspect Search Engine User Behavior Study, *http://www.iprospect.com/about/whitepaper_seuserbehavior_apr06.html*, (2006).

[10] C. de Kerchove, L. Ninove, Maximizing PageRank via outlinks, *Lin. Alg. Appl.*, 429(2008): 1254-1276.

[11] A. N. Langville, C. D. Meyer, The use of linear algebra by Web search engines, *IMAGE Newsletter*, 33 (2004) 2-6.

[12] A. N. Langville, C. D. Meyer, Deeper inside PageRank, *Internet Mathematics*, 1(3) (2005) 335-380.

[13] N. Ma, J. Ch. Guan, Y. Zhao, Bringing PageRank to the citation analysis, *Information Processing & Management*, 44(2008): 800-810.

[14] G. R. Notess, Rising relevance in search engines, *Online1999*, 23(3): 84-86.

[15] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, Detecting spam web pages through content analysis, *In Proc. 15th WWW*, (2006): 83-92.

[16] L. Page, S. Brin, The PageRank citation ranking: Bringing order to the web, *http://google.stanford.edu/ backrub/pageranksub.ps*, (1998)

[17] M. Richardson, P. Domnigos, The intelligent surfer: Probabilistic combination of link and content information in PageRank, *Advances in Neural Information Processing Systems*, 14(2002): 637-680

[18] K. Sugiyama, et al., Improvement in TF-IDF Scheme for Web Pages Based on the Contents of Their Hyperlinked Neighboring Pages, *Syst. Comp. Jpn*, 36(14)(2005): 56-68.

[19] A. Sidi, Vector extrapolation methods with applications to solution of large systems of equations and to PageRank computations, *Comput. Math. Appl.*, 56(2008): 1-24.

[20] H. Sun, Y. M. Wei, A Note on the PageRank algorithm, *Appl. Math. Comp.*, 179(2006): 799-806.

[21] E. Ward, Market through 'link analysis' to improve, popularity quality. *Advertising Age's Business Marketing*, 85(2000) 32-33.

[22] W. Xing, A. Ghorbani, Weighted PageRank Algorithm, *Proceedingd of the 2nd Annual Conference on IEEE Communication Networks an Service Research*, (2004)