ISSN:0975-9646

Harjeet Kaur et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3085 - 3088

Impact of PageRank and its Variation On Convergence and Ranking

Harjeet Kaur[1] and Dr. Kanwal Garg[2]

Scholar(M.Tech CSE)[1] , Assistant Professor[2]
Department of Computer Science and Applications,
Kurukshetra University, Kurukshetra India

Abstract: Page importance information has a direct influence on web search effectiveness because if a useful page gets lower page importance value, it will be absent or ranked very low in search results. Still there has been little work on measuring this effect. In this paper, the author observed mechanical method of calculating page importance through PageRank formula. The standard PageRank formula is then compared with its two variations named weighted PageRank(WPR) and TrustRank(TR). The Weighted Personalised PageRank(WPPR), a combination of WPR and TR, is introduced in this paper. The mathematically calculated results show that WPPR performs faster and efficient than its ancestors in terms of retrieving more relevant pages earlier.

Keywords: Standard PageRank, Web Graph, Random Link, Convergence

1. INTRODUCTION:

Search Engines are the primary discovery mechanism for pages on the web and the web has an infinite set of pages. Search engines use crawlers that extract links, starting from a set of seeds, to discover new pages. However not all the pages are of equal importance.

Computing page importance is a valuable aspect of crawling as search engines display results in the order of page importance. It is also helpful in discovery of new pages as important pages are to be fetched first. Beside this, it also aid refreshing policy of web pages as important pages should be refreshed more often[4].

Among various page importance metric, link based metrics provide an objective measure of page importance that corresponds well with people's subjective idea of importance[5]. PageRank is the first and successful metric in this category. It is a good measure of page quality, better than just counting the number of in-links[1]. However the other criteria such as occurrences of words from query, their position, user interests are not the focus of this paper.

Although well-known algorithms exist for ranking page importance, the experiments comparing their effectiveness are hardly published. The hindrance in this path is the infinite size and dynamicity of web. In this paper, a new variation is proposed on the intuition that a formula considering the importance of links based on the popularity as well as the trust score of that page will be able to explore the highly useful pages earlier, when the formulas based on single criteria individually are performing better than standard PageRank formula.

The rest of the paper is organized as follows: Section 2 provides references to related work in this area. A comparative analysis of PR and its variations is proposed in section 3.A new variation is proposed in section 4 which is then analyzed in section 5. Finally section 6 summarizes the conclusion and future work.

2. RELATED WORK

Page and Brin were the first to introduce the idea of page importance based on link structure of web, The PageRank was proposed as a model of user's success of Google search engine which implemented PageRank proved the significance of PageRank. In [5] the importance of web pages was based on a PageRank metric. It state that if a page has important links to it , its link to other pages also contribute to their importance and a page with high PageRank is most relevant page to be downloaded. The PageRank of a page A is given by:

$$PR(A)= (1-d)/|D| + d(PR(T1)/C(T1)+….+PR(Tn)/C(Tn))$$

Where PR (A) = PageRank of page A

T1....Tn= inlinks to page A

C(A) = no. of links going out of page A

D = set of all web pages

d= damping factor which is often assumed to 0.85.

This metric measures the importance of a page very effectively as shown by [3][5] but require multiple

calculations over a large web graph and can be easily spammed. The two important variations of PageRank proposed in literature are:

2.1 Weighted PageRank: In [6] an improved version of PageRank which assigns more value to more important pages instead of dividing the rank value of a page evenly among its entire outgoing links. The weighted PageRank is thus given by:

PR(u) = (1-d)/|D| + d(PR(V1)Win(V1,u)Wout(V1,u)+….+PR(Vn)Win(Vn,u) Wout(Vn,u)

Where Win(v,u) = weight of link(v,u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v

Wout(v,u) = weight of link(v,u) calculated based on the number of outlinks of page u and the number of inlinks of all reference pages of page v

PR(u)=PageRank of page u

2.2 Personalised PageRank [7]: To determine the deserved ranking of web pages [7] proposed another variation of Standard Pagerank that aid web spam detection by assuming that a user goes to a trusted site rather than to every page with equal probability. So the PageRank of page A is then defined as

PR(Ai)= (1-d)(Ti) + d(PR(R1)/C(R1)+….+PR(Rn)/C(Rn))

Where PR (A) = PageRank of page A

R1....Rn= inlinks to page A

Ti = trust score of page i

C(A) = no. of links going out of page A

d= damping factor which is often assumed to 0.85

As users are unlikely to go a single trusted page a new form named windowed rankmass algorithm was adapted in [2]. This new form batches together sets of probability calculation and downloading sets of pages at a time thereby reducing the computational overhead.

3. COMPARATIVE ANALYSIS OF PR AND ITS VARIATIONS:

To evaluate the convergence nature of PageRank formula, we observed the PageRank value at different iterations for two different initial values- first being the smallest, say 0 and the second being the largest, say 40. The outcome of this observation is that the convergence of formula to its approximate value is quite faster in first 10-15 iterations.So it is quite interesting to study the convergence properties of various page importance calculating formulas.

In order to study the difference between PageRank and its variation in terms of computation cost and importance order, initially we considered a 2 page graph .We assumed, for simplicity, Page A as good page having trust score=1 and Page B as bad page having trust score=0.The calculation started with initial PageRank value of Page B = 0.

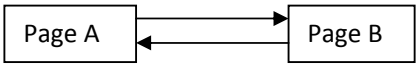

Figure 1 : A two page hypothetical graph

Both PR and WPR performs equally in this case as links are quite simple. They both are outperformed by TrustRank formula which is faster to converge and requires less iterations. It is therefore less expensive computationally for 2 page graph. Another important observation is that precision does not increased often faster in TR as against PR and WPR that works towards increasing precision continuously.

So before coming to any conclusion, another graph of 3 pages is taken under consideration.

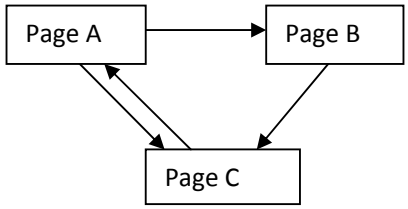

Figure 2: Hypothetical strongly connected graph

For this graph, we considered two versions of TR formula, TR(i) and TR(ii) that are based on the different trust score assignment. The technique to assess trust score of each web page is explained in [7].In this paper for ease, we considered an ignorant trust function proposed in [7] which is described as follows:

T(p) = 0 if P is a good page
1 if P is a bad page
½ otherwise
Where T(p) = Trust score of Page P

In the present work, the trust score assignment used is given in table 1.

	Page A	Page B	Page c
TR(i)	1	0	½
TR(ii)	½	0	1
WPPR(i)	1	0	½
WPPR(ii)	1/2	0	1

Table:1 Trust Score Values

In this case, WPR performs well by converging in just 7 iterations. PageRank is behind WPR by converging in 9 iterations. The TR is quite slower and took 12-13 iterations to converge to a suitable value as shown in figure 3.

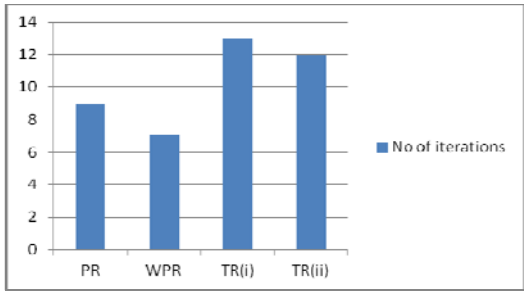

Figure 3: Iterations Required for convergence for 3 page graph

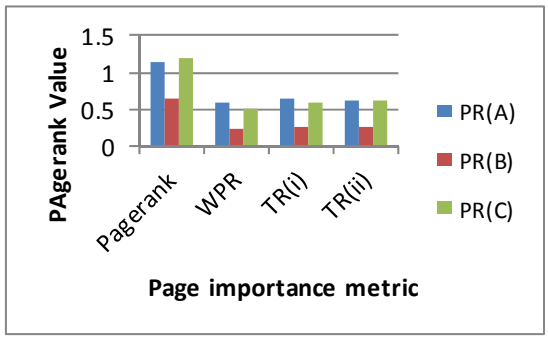

Figure 4 : Order of importance through various metrics

The order of importance of each page exhibited by different PageRank formulas is shown in figure 4. The PR assign highest value to page with large number of inlinks whereas TR assign higher value to more trusted page. The WPR provided different page ranking depending upon the popularity of links.

4. Proposed Variation

The above two variations namely WPR and TR gives an insight of hybrid formula that distribute rank score according to popularity of links as well as ensures the crawling of only trusted pages.

The proposed formula thus have the following form:

$Pr(b)= d \sum_{a \epsilon B(b)} (PR(a)\ Win(a,b)\ Wout(a,b)) +(1-d)\ ti$

Where d= dampening factor

Ti=trust score of page i

Win(v,u) = weight of link(v,u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v

Wout(v,u) = weight of link(v,u) calculated based on the number of outlinks of page u and the number of inlinks of all reference pages of page v

5. Result Analysis

The proposed formula WPPR works more efficiently by converging in 7-8 iterations. It is thus faster and less expensive computationally. Also it incorporates the best part of two already proved effective variations of PR. The precision also increases quite faster after two place convergence.

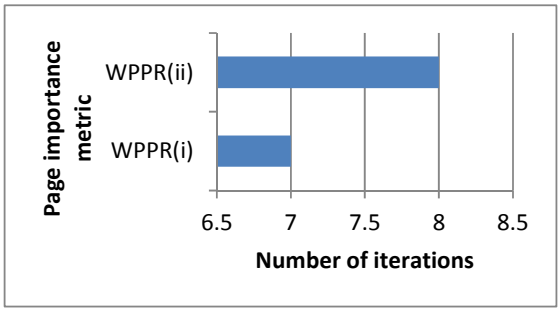

Figure 5: Convergence of Weighted Personalised PageRank for 3 page graph

The following figure gives the order of importance by the proposed formula.

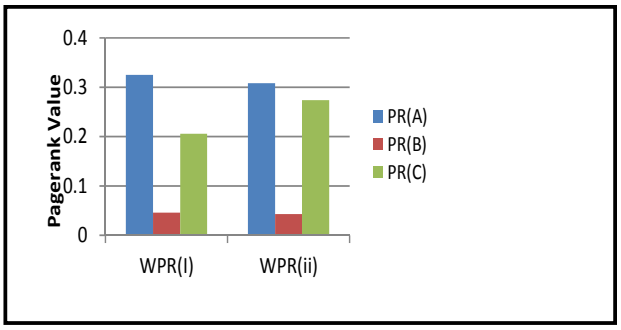

Figure 6 : Order of importance through weighted personalised Page Rank

6. CONCLUSION AND FUTURE SCOPE

Most of the page importances metrics present in literature were able to download important pages first. In this paper the original PageRank formula is observed to determine the convergence and precision property. Other variations named WPR and TR are also observed for same hypothetical graphs along with standard PageRank. All the three formulas were important to yield good search results. However WPR is faster to converge and provide higher precision than the other two. The two variations studied, of conventional PR, allowed us to propose another hybrid PageRank formula named as Weighted Personalised PageRank. The newly introduced variation provided more relevant results than the other two by converging earlier and providing higher precision in later convergence.

However the implementation of proposed formula in an algorithm is still a future work. Also, determining its usage on larger web graph is a great research avenue. In addition, there are number of ways to refine our formula. For instance, the allotment of trust score based on weights of links.

REFERENCES

[1] Dennis Fetterly, Nick Craswell, Vishwa Vinay, The impact of Crawl Policy On web Search Effectiveness, In Proceedings of SIGIR, July 2009.

[2] J. Cho and U. Schonfeld.Rankmass crawler: a crawler with high personalized PageRank coverage guarantee. In Proceedings of VLDB, pages 375–386, 2007.

[3] Lawrence Page, Sergery Brin, Rajeev Motwani and Terry Winograd, The PageRank Citation Ranking, Bringing order to the web, 1998.

[4] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In WWW '03: Proceedings of the 12th international conference on World Wide Web, New York USA, pages 280–290, 2003.

[5] S. Brin and L. Page.The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, 30(1{7):107{117, April 1998.

[6] Wenpu Xing and Ghorbani Ali, Weighted PageRank algorithm, In Proceeding of the second annual conference on Communication Networks and Services Research(CSNR' 04),IEEE,2004.

[7] Z. Gyongyi, H. Garcia-Molina and J. Pederson. Combating Web Spam with Trustrank in Proceeding of the 30th VLDB conference, Toronto, Canada, 2004.