Weighted PageRank using the Rank Improvement

Rashmi Rani*, Vinod Jain**

* B.S.Anangpuria. Institute of Technology and Management, Faridabad, India
** Department of Computer Science and Engineering B.S.Anangpuria. Institute of Technology and Management, Faridabad, India

Abstract- Information available on the WWW, users' get easily lost in rich hyper structure. It has become increasingly necessary for user's to utilize automated tool in order to find, extract, filter and evaluate the desired information and resources. Modern Information Retrieval System matches the term of a user with documents in their index and returns a large number of documents of Web pages generally in the form of ranked list. It becomes almost impractical at the user end to examine every return document, thus the need to look for some result optimization. In this paper, finding the content of the web and retrieving the users' interest and need from their behaviour has become important .Web mining is used to cauterized user and pages by analyzing the user's behaviour, content of pages, order of the URLs, Two page ranking algorithms, HITS and PageRank. The Weighted Page Rank algorithm (WPR), takes into account the importance of both inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages. In this paper we have the relevancy values for the query produced by Page Rank and WPR using different page set and finally search result list is reranked by updating the existing page rank values of a page. The proposed work result WPR performs better than Page Rank and reduced search time and important pages are tending to move upwards in the result list.

Index Terms- Page Rank algorithm, Weighted Page Rank, Web Mining, Web Structure Mining, HITS

I. INTRODUCTION

WWW is one of the popular information resources for text, image, audio, video, and metadata. It is estimated that WWW has expanded by about 2000 % since its inception and is doubling in size every six to ten months [1].Nowadays providing a set of web pages based on user query words is not a big problem in search engine[2],instead the problem is that a search engine return a large number of web pages in response to user queries and user have to spend much time in finding desired information from this long list resulting in information overload problem[3].although many search engine provide a user friendly ranked list in response to user queries, there still remain a challenge in finding the desired information content within the result list due to the presence of user desired pages being scattered throughout in the result list.

Web mining is used to discover the content of the Web, the users' behaviour in the past, and the WebPages that the users want to view in the future. Web mining consists of Web Content Mining (WCM), Web Structure Mining (WSM), and Web Usage Mining (WUM) [4, 5, 6]. WCM deals with the discovery of useful information from web content.WSM discovers relationships between web pages by analyzing web structures. WUM ascertains user profiles and the users' behaviour recorded inside the web logfile.WCM focuses mainly on the structure within a document (the inner-document level) while WSM tries to discover the link structure of the hyperlinks between documents (the inter document level). The numbers of inlinks (links to a page) and of outlinks (links from a page) are valuable information in web mining. This is due to the facts that a popular webpage is often referred to by other pages and that an "important" webpage contains a high number of outlinks. Therefore, WSM is seen as an important approach to web mining [7]. This paper focuses on WSM and provides a new Weighted Page Rank Algorithm and reranked by updating the existing Page Rank values.

The work proposed in this paper is aim to optimize the result of a search engine by returning the more relevant and user desired page on top of search result list. To perform the required task, the work take into account the mined output of WSM. The paper has been organised as follows: Section II describe the background terminology used in the work; Section III present the PageRank algorithm which is used by WSM. Section IV describes the extended PageRank algorithm called the Weighted PageRank algorithm. Section V describes the different component involved in the implementation and evaluation of WPR. Section VI the experimental result and their implication for WPR .Section VII summarizes and conclusion .Finally the result re-rank of PageRank and WPR for query are given in Appendices A and B respectively.

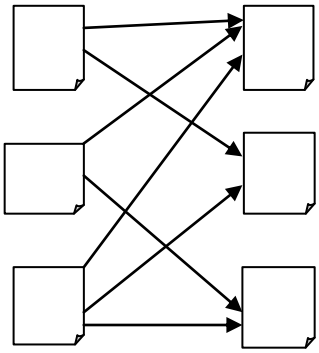
II. RELATED BACKGROUND TERMINOLOGY

Most of the search engines use *Page ranking* algorithms, which can arrange the documents in order of their relevance, importance and content score. Some search engines also apply Web Mining techniques such as clustering, classification, association rule discovery and categorization to filter, classify as well as group their search results. Many page ranking algorithms [8, 9] have been proposed in the literature such as *HITS, Clever, PageRank, Weighted PageRank, Page Content Rank.* Some algorithms rely only on the link structure of the documents i.e. their popularity scores (web structure mining), some look for the content of the documents with respect to the user query (web content mining), while others use a combination of both i.e. they use links as well as the content of the document to assign a rank value to the concerned document.

Google first retrieves a list of relevant pages to a given query based on factors such as title tags and keywords. Then it uses PageRank to adjust the results so that more "important" pages are provided at the top of the page list.

HITS ranks WebPages by analyzing their inlinks and outlinks. In this algorithm, WebPages pointed to by many h

hyperlinks are called *authorities* whereas WebPages that point to many hyperlinks are called *hubs* [10, 11, 12]. Authorities and hubs are illustrated in Figure 1.



Hubs Authorities

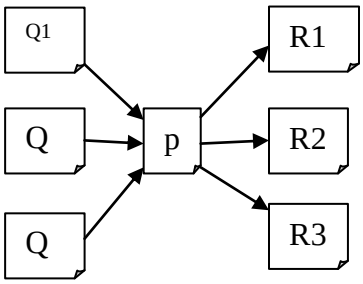
Figure 1. Hubs and authorities

Hubs and authorities are assigned respective scores. Scores are computed in a mutually reinforcing way: an authority pointed to by several highly scored hubs should be a strong authority while a hub that points to several highly scored authorities should be a popular hub [9, 10]. Let ap and hp represent the authority and hub scores of page p, respectively. B (p) and I (p) denote the set of referrer and reference pages of page p, respectively. The scores of hubs and authorities are calculated as follows [9, 10, 13]:

$$ap = \sum_{q \in B(p)} hq \quad (1)$$

$$hp = \sum_{q \varepsilon I(p)}^{n} aq \quad (2)$$

Figure 2 shows an example of the calculation of authority and hub scores.



ap = hq1 + hq2 + hq3 hp = ar1 + ar2 + ar3

Figure 2. An example of HITS operations

HITS is a purely link-based algorithm. It is used to rank pages that are retrieved from the Web, based on their textual contents to a given query. Once these pages have been assembled, the HITS algorithm ignores textual content and focuses itself on the structure of the Web only.

III. PAGERANK ALGORITHM

Page Rank [14, 15, 16] was developed at Stanford University by Larry page (cofounder of Google search engine) and Sergey Brin. Google uses this algorithm to order its search results in such a way that important documents move up in the results of a search while moving the less important pages down in its list. This algorithm states that if a page has some important incoming links to it, then its outgoing links to other pages also become important, thus it takes backlinks into account and propagates the ranking through links. When some query is given, Google combines precomputed PageRank scores with text matching scores to obtain an overall ranking score for each resulted web page in response to the query. Although many factors determine the ranking of Google search results but PageRank continues to provide the basis for all of Google's web search tools.

A simplified version of PageRank is defined in (3):

$$PR(u) = c \sum_{V \varepsilon B(u)} \left(\frac{PR(v)}{Nv} \right) \quad (3)$$

Where *u* represents a web page, *B* (*u*) is the set of pages that point to *u*. *PR* (*u*) and *PR* (*v*) are rank scores of page *u* and *v*, respectively. *Nv denotes* the number of outgoing links of page *v*, *c* is a factor used for normalization.

In PageRank, rank score of a page *p* is evenly divided among outgoing links. Values assigned to the outgoing links of page *p*are in turn used to calculate the ranks of the pages to which page *p* is pointing. Example showing distribution and assignment of page ranks is illustrated in Figure 3.

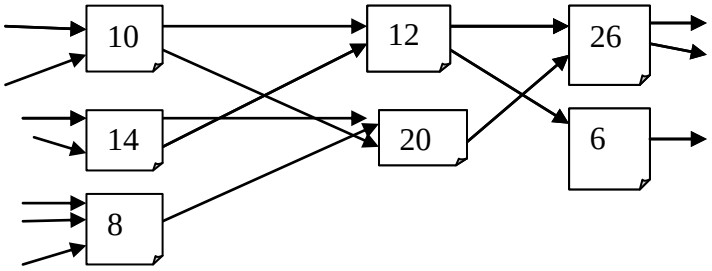


Figure 3. Distribution of PageRank

Later PageRank was modified keeping in view the Random Surfer Model [16] which states that not all users follow the direct links on WWW. The modified version is given in (2).

$$PR(u) = (1 - d) + d \sum_{V \varepsilon B(u)} \left(\frac{PR(v)}{Nv} \right) \quad (5)$$

Where *d* is a damping factor [6] that is usually set to 0.85. *d* can be thought of as the probability of users‟ following the direct

links and $(1 - d)$ as the page rank distribution from non-directly linked pages.

IV. Weighted PageRank Algorithm

Wenpu Xing and Ali Ghorbani [7] proposed an extension to standard PageRank called *Weighted PageRank (WPR)*. It assumes that more popular the web pages are more linkages other web pages tend to have to them or are linked to by them. This algorithm assigns larger rank values to more important pages instead of dividing the rank value of a page evenly among its outgoing linked pages. Each outlink page gets a value proportional to its popularity or importance and this popularity is measured by its number of incoming and outgoing links. The popularity is assigned in terms of weight values to the incoming and outgoing links, which are denoted as *Win*(*v*,*u*) and *Wout*(*v*,*u*) respectively. *Win* (*v*, *u*) (given in (3)) is the weight of *link* (*v*, *u*) calculated based on the number of incoming links of page *u* and the number of incoming links of all reference (outgoing linked) pages of p.

$$W^{in}(v,u) = \frac{Lu}{\sum_{p \varepsilon R(v)} Lp} \quad (6)$$

Where *Iu* and *Ip* represent the number of inlinks of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*. *Wout* (*v*, *u*) (given in (4)) is the weight of *link* (*v*, *u*) calculated based on the number of outlinks of page *u* and the number of outlinks of all reference pages of page *v*.

$$W^{out}(v,u) = \frac{Ou}{\sum_{p \varepsilon R(v)} Op} \quad (7)$$

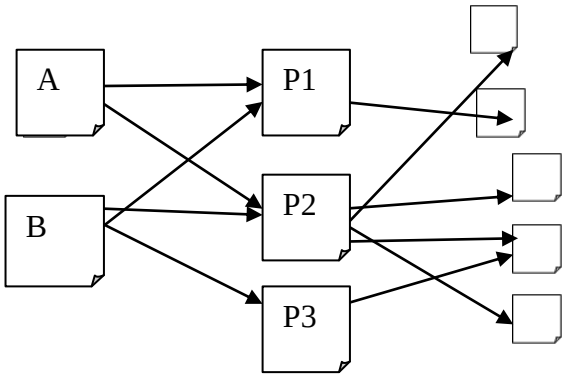


Figure 5. Links of a website

In this example, Page A has two reference pages: p1 and p2. The inlinks and outlinks of these two pages are Ip1 = 2, Ip2 = 1, Op1 = 2, and Op2 = 3. Therefore,

$$W^{out}(A,P,I) = \frac{Op1}{Op1+Op2} = \frac{2}{5} \quad (8)$$

Considering the importance of pages, the original PageRank formula is modified as

$$PR(u) = (1-d) + d \sum_{V \varepsilon B(u)} PR(v)\, W^{in}(v,u)\, W^{out}(v,u) \quad (9)$$

V. Experiments

To evaluate the WPR algorithm, we implemented WPR and the standard PageRank algorithms to compare their results. Figure 6 illustrates different components involved in the implementation and evaluation of the WPR algorithm. The simulation studies we have carried out in this work consist of six major activities:

1. *Finding a web site*: Finding a web site with rich hyperlinks is necessary because the standard PageRank and the WPR algorithms rely on the web structure. After comparing the structures of several web sites, the website Of Saint Thomas University, in Fredericton, has been chosen.
2. *Building a web map*: There is no web map available for this website. A free spider software–J Spider–is used to generate the required web map.
3. *Finding the root set*: A set of pages relevant to a given query is retrieved using the IR search engine embedded in the web site. This set of pages is called the *root set*.
4. *Finding the base set*: A base set is created by expanding the root set with pages that directly point to or are pointed to by the pages in the root set.
5. *Applying algorithms*: The Standard PageRank and the WPR algorithms are applied to the base set.
6. *Improve the weighted rank*
7. *Evaluating the results*: The algorithms are evaluated by comparing their results.

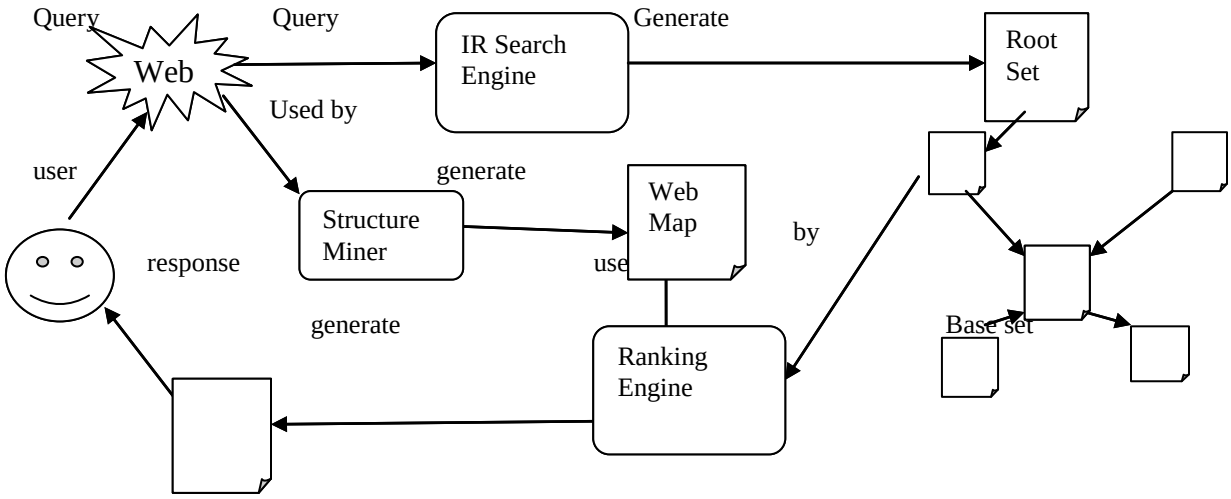


Figure 6. Architectural components of the system used to implement and evaluate the WPR algorithm

VI. EVALUATION

The query topics "travel agent" and "scholarship" are used in the evaluation of the WPR and the standard Page Rank algorithms. "Travel agent" represents a non-focused topic whereas "scholarship" represents a focused (popular) topic in the Website of Saint Thomas University. The results of the evaluation are summarized in the following subsections.

6.1. The determination of the relevancy of the pages to the given query

The Standard Page Rank and the WPR algorithms provide important information about a given query by using the structure of the website. We categorized the pages in the results into four classes based on their relevancy to the given query:

- **Very Relevant pages (VR)**, which contain very important information about the given query,
- **Relevant pages (R)**, which have relevant but not important information about the given query,
- **Weak-Relevant pages (WR)**, which do not have relevant information about the given query even though they contain the keywords of the given query, and
- **Irrelevant pages (IR)**, which include neither the keywords of the given query nor relevant information about it.

6.2. The Calculation of the relevancy of the page lists to the given query

The performances of the WPR and the standard Page Rank algorithms have been evaluated to identify the algorithm that produces better results

Table 1. The relevancy values for the query "travel agent" produced by PageRank and WPR using different page sets

	Number of Relevant Pages		Relevancy Value(k)	
Size of page set	**PageRank**	**WPR**	**PageRank**	**WPR**
10	0	1	0.1	0.5
20	4	3	13.1	16.8
30	4	4	47.1	49.8
40	4	4	82.1	84.8
50	4	4	117.1	119.8
60	5	5	159.6	162.3
70	7	7	211.7	214.4

6.3 Rank Updator

This module takes in input from the query Rank improvement: This module takes the input from the query processor and matched documents of a user query and an improvement is applied to improve the rank score of the returned pages. The module operates online at the query time and applied the improvement on the concerned documents.

Step 1: Given an input user query q and matched document D collected from the query processor, the web page is found to which the query q belongs.

Step 2: The level weight are calculated for every page X present in the sequential pattern.

Step 3: The rank are calculated for every page X present in the sequential pattern. The improved is calculated as the summation of pervious rank and assigned weight value.

By improving the rank, the result of a search engine can be optimization so as to better serve the user need. The user can now find the popular and relevant pages upwards in the result list.

Rank Improvement

The rank of a page can be improved with the help of its assign weight. The new rank can become:

New_Rank(X) = Rank(X) + Weight (11)

Where rank(X) is the existing rank value (PageRank) of page X WPR(X) is the Popularity given to X.

Relevancy Rule: the relevancy of a page to a given query depends on its category and its position in the page-list. The larger the relevancy value is, the better is the result. The relevancy, _, of a page-list is a function of its category and position:

$$X = \sum_{i \varepsilon R(p)} (n - i) * Wi \quad (12)$$

where i denotes the ith page in the result page-list R(p), n represents the first n pages chosen from the list R(p), and Wi is the weight of page i.

Wi = V1 ,if the ith page in PR
V2, if the ith page in R
V3, if the ith page in WR
V4, if the ith page in IR

Where v1>v2>v3>v4

VII. Experimental Result

	Number of Relevant Pages		Relevancy Value(X)		New Relevancy Value
Size of the paper set	**PageRank**	**WPR**	**PageRank**	**WPR**	**(PageRank + WPR)**
10	0	1	0.1	0.5	0.6
20	4	3	13.1	16.8	29.9
30	4	4	47.1	49.8	96.9
40	4	4	82.1	84.8	166.9
50	4	4	117.1	119.8	236.9
60	5	5	159.6	162.3	321.9

VIII. Conclusion

Web mining is used to extract information from users' past behavior. Web structure mining plays an important role in this approach. Two commonly used algorithms in web structure mining are HITS and PageRank, which are used to rank the relevant pages. Both algorithms treat all links equally when distributing rank scores. Several algorithms have been developed to improve the performance of these methods. This paper introduces the WPR algorithm, an extension to the PageRank algorithm. WPR takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages. Saint Thomas University show that WPR is able to identify a larger number of relevant pages to a given query compared to standard PageRank.

Web mining is used to extract useful information from Users' past behavior. In this paper the Page Rank and Weighted Page Rank algorithms are used by many search engine but the users may not get the required relevant documents easily on the top few pages. To solve this problem we use the Weighted Page Content Rank has been proposed which which employ Web structure mining as well as Web Content mining technique. This algo is improving the order of the page in the result list so that the user gets the relevant and important pages in the list.

References

[1] NareshBarsagade,"Web usage mining and pattern discovery: A survey paper".CSE8331, Dec, 2003.

[2] A.Arasu, J. Cho, H. Garcia-Molina,A. Paepcke, and S.Raghavan ,"Searching the Web "[3] A. Brochers, J. Herloker, J. Konstanand, and J.Riedl,"Ganging up on information overload," Computer,Vol.31,No.4,pp.106-108,1998.

[3] R. Kosala and H. Blockeel. Web mining research: A survey.*ACM SIGKDD Explorations*, 2(1):1–15, 2000.

[4] S. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim. Research issues in web data mining. In *Proceedings of the* Conference on Data Warehousing and Knowledge Discovery, pages 303–319, 1999.

[5] S. Pal, V. Talwar, and P. Mitra. Web mining in soft computing framework: Relevance, state of the art and future directions. *IEEE Trans. Neural Networks*, 13(5):1163–1177, 2002.

[6] Wenpu Xing and Ali Ghorbani," Weighted PageRank Algorithm" *Faculty of Computer Science University of New Brunswick Fredericton, NB, E3B 5A3, Canada* E-mail: {m0yac, ghorbani}@unb.ca

[7] NeelamDuhan, A. K. Sharma, Komal Kumar Bhatia, "Page Ranking Algorithms: A Survey". In proceedings of the IEEE International Advanced Computing Conference (IACC), 2009.

[8] JaroslavPokorny, JozefSmizansky, "Page Content Rank: An approach to the Web Content Mining".

[9] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon. Link analysis: Hubs and authorities on the world. *Technical report: 47847*, 2001.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment.*Journal of the ACM*, 46(5):604–632, September 1999.

[11] J. Wang, Z. Chen, L. Tao, W. Ma, and W. Liu.Ranking user's relevance to a topic through link analysis on web logs. *WIDM*, pages 49–54, 2002.

[12] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg.

[13] L. Page, S. Brin, R. Motwani, T. Winograd, "The page rank citation ranking: Bringing order to the web". Technical report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[14] C. Ridings and M. Shishigin, "Page rank uncovered". Technical report, 2002.Mining the Web's link structure. *Computer*, 32(8):60–67, 1999.

[15] http://pr.efactory.de/e-pagerank-algorithm.shtml

Authors

First Author – Rashmi Rani, B.S.Anangpuria. Institute of Technology and Management, Faridabad, India
Second Author – Vinod Jain, Department of Computer Science and Engineering B.S.Anangpuria. Institute of Technology and Management, Faridabad, India