Weighted PageRank Algorithm

Wenpu Xing and Ali Ghorbani
Faculty of Computer Science
University of New Brunswick
Fredericton, NB, E3B 5A3, Canada

E-mail: {m0yac,ghorbani}@unb.ca

Abstract

With the rapid growth of the Web, users get easily lost in the rich hyper structure. Providing relevant information to the users to cater to their needs is the primary goal of website owners. Therefore, finding the content of the Web and retrieving the users' interests and needs from their behavior have become increasingly important. Web mining is used to categorize users and pages by analyzing the users' behavior, the content of the pages, and the order of the URLs that tend to be accessed in order. Web structure mining plays an important role in this approach. Two page ranking algorithms, HITS and PageRank, are commonly used in web structure mining. Both algorithms treat all links equally when distributing rank scores. Several algorithms have been developed to improve the performance of these methods. The Weighted PageRank algorithm (WPR), an extension to the standard PageRank algorithm, is introduced in this paper. 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. The results of our simulation studies show that WPR performs better than the conventional PageRank algorithm in terms of returning larger number of relevant pages to a given query.

Keywords: Web Mining, Web Structure Mining, HITS, PageRank, Weighted PageRank

1. Introduction

In the highly competitive world and with the broad use of the Web in e-commerce, e-learning, and e-news, finding users' needs and providing useful information are the primary goals of website owners. Therefore, analyzing users' patterns of behavior becomes increasingly important.

Web mining is used to discover the content of the Web, the users' behavior 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) [6, 7, 9]. 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' behavior recorded inside the web logfile. WCM and WUM have been studied by many researchers who have achieved valuable results. Based on the topology of the hyperlinks, WSM categorizes web pages and generates related patterns, such as the similarity and the relationships between different Web sites. Technically, 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. This paper focuses on WSM and provides a new Weighted PageRank Algorithm.

The rest of this paper is organized as follows. A brief background review of web structure mining is presented in the next section. Section 3 presents the PageRank algorithm, a commonly used algorithm in WSM. An extended PageRank algorithm called the Weighted PageRank algorithm (WPR) is described in Section 4. Different components involved in the implementation and evaluation of WPR are presented in Section 5. The experimental results and their implication for WPR are given in Section 6. Section 7 summarizes the conclusions of the present study. Finally, the result sets of PageRank and WPR for the query "travel agent" are given in Appendices A and B respectively.

Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04)
0-7695-2096-0/04 $20.00 © 2004 IEEE

2. Background

With the rapid growth of the Web, providing relevant pages of the highest quality to the users based on their queries becomes increasingly difficult. The reasons are that some web pages are not self-descriptive and that some links exist purely for navigational purposes. Therefore, finding appropriate pages through a search engine that relies on web contents or makes use of hyperlink information is very difficult.

To address the problems mentioned above, several algorithms have been proposed. Among them are PageRank [10] and *Hypertext Induced Topic Selection* (HITS) [2, 9] algorithms. PageRank is a commonly used algorithm in Web Structure Mining. It measures the importance of the pages by analyzing the links [1, 8]. PageRank has been developed by Google and is named after Larry Page, Google's co-founder and president[10]. PageRank ranks pages based on the web structure.

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 [10]. The Pagerank algorithm is described in detail in the next section.

HITS ranks webpages by analyzing their inlinks and outlinks. In this algorithm, webpages pointed to by many hyperlinks are called *authorities* whereas webpages that point to many hyperlinks are called *hubs* [4, 5, 11]. Authorities and hubs are illustrated in Figure 1.

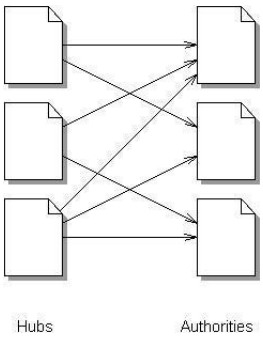

Figure 1. Hubs and authorities

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 [4, 5]. Let a_p and h_p 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 [2, 4, 5]:

$$a_p = \sum_{q \in B(p)} h_q \quad (1)$$

$$h_p = \sum_{q \in I(p)} a_q \quad (2)$$

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

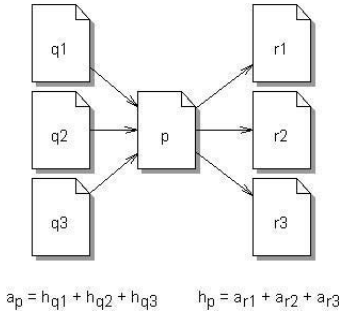

Figure 2. An example of HITS operations

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. Some difficulties arise from this feature [2]:

- HITS frequently returns more general webpages on an otherwise narrowly focused topic because the web does not contain many resources for the topic,
- Topic drift occurs while the hub has multiple topics because all of the outlinks of a hub page get equivalent weights, and
- Some popular sites that are not highly relevant to the given query gain overhead weight values.

The CLEVER algorithm is an extension of standard HITS and provides an appropriate solution to the problems that result from standard HITS [2]. CLEVER assigns a weight to each link based on the terms of the queries and end-points of the link. It combines anchor text to set weights to the links as well. Moreover, it breaks large hub pages into smaller units so that each hub page is focused on as a single topic. Finally, in the case of a large number of pages from a single domain, it scales down the weights of pages to reduce the probabilities of overhead weights [2].

Another major shortcoming of standard HITS is that it assumes that all links pointing to a page are of equal

weight and fails to recognize that some links might be more important than others. A *Probabilistic analogue of the HITS Algorithm*(PHITS) has been developed to solve this problem[3]. PHITS provides a probabilistic interpretation of term-document relationships and identifies authoritative documents. In the experiment on a set of hyperlinked documents, PHITS demonstrates better results compared to those obtained by standard HITS. The most important feature of the PHITS algorithm is its ability to estimate the actual probabilities of authorities compared to the scalar magnitudes of authority that are provided by standard HITS[3].

3. The PageRank Algorithm

The PageRank algorithm, one of the most widely used page ranking algorithms, states that if a page has important links to it, its links to other pages also become important. Therefore, PageRank takes the backlinks into account and propagates the ranking through links: a page has a high rank if the sum of the ranks of its backlinks is high [8, 10]. Figure 3 shows an example of backlinks: page A is a backlink of page B and page C while page B and page C are backlinks of page D.

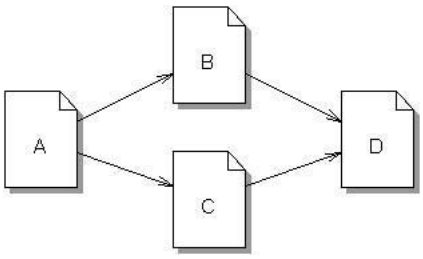

Figure 3. An example of backlinks

3.1. Simplified PageRank

A slightly simplified version of PageRank is defined as [8]:

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{N_v} \tag{3}$$

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. N_v denotes the number of outgoing links of page v. c is a factor used for normalization. Figure 4 shows an example in which $c = 1.0$ to simplify the calculation.

In PageRank, the rank score of a page, p, is evenly divided among its outgoing links. The 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. The rank scores of pages of a website could be calculated iteratively starting from any webpage. Within a website, two or more pages might connect to each other to form a loop. If these pages did not refer to but are referred to by other webpages outside the loop, they would accumulate rank but never distribute any rank. This scenario is called a *rank sink* [8].

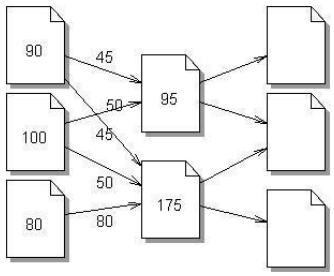

Figure 4. An example of simplified version of PageRank

3.2. PageRank

To solve the *rank sink* problem, we observed the users' activities. A phenomenon is found that not all users follow the existing links. For example, after viewing page a, some users may not decide to follow the existing links but directly go to page b, which is not directly linked to page a. For this purpose, the users just type the URL of page b into the URL text field and jump to page b directly. In this case, the rank of page b should be affected by page a even though these two pages are not directly connected. Therefore, there is no absolute *rank sink*.

Based on the consideration of the phenomenon mentioned above, the original PageRank is published [8, 10]:

$$PR(u) = (1 - d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \tag{4}$$

where d is a dampening factor that is usually set to 0.85. We also could think of d as the probability of users' following the links and could regard $(1 - d)$ as the pagerank distribution from non-directly linked pages.

To test the utility of the PageRank algorithm, Google applied it to the Google search engine [8]. In the experiments, the PageRank algorithm works efficiently and effectively because the rank value converges to a reasonable tolerance in the roughly logarithmic $(\log n)$ [8, 10].

The rank score of a web page is divided evenly over the pages to which it links. Even though the PageRank algorithm is used successfully in Google, one problem still ex-

ists: in the actual web, some links in a web page may be more important than are the others.

4. Weighted PageRank (WPR)

The more popular webpages are, the more linkages that other webpages tend to have to them or are linked to by them. The proposed extended PageRank algorithm–a Weighted PageRank Algorithm–assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$, respectively.

$W^{in}_{(v,u)}$ is the 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.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p} \tag{5}$$

where I_u and I_p represent the number of inlinks of page u and page p, respectively. $R(v)$ denotes the reference page list of page v.

$W^{out}_{(v,u)}$ 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.

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} O_p} \tag{6}$$

where O_u and O_p represent the number of outlinks of page u and page p, respectively. $R(v)$ denotes the reference page list of page v.

Figure 5 shows an example of some links of a hypothetical website.

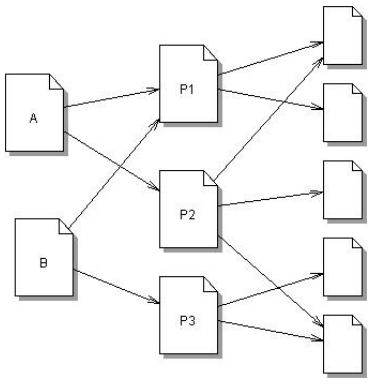

Figure 5. Links of a website

In this example, Page A has two reference pages: $p1$ and $p2$. The inlinks and outlinks of these two pages are $I_{p1} = 2$, I_{p2} = 1, O_{p1} = 2, and O_{p2} = 3. Therefore,

$$W^{in}_{(A,p1)} = I_{p1}/(I_{p1} + I_{p2}) = \frac{2}{3}$$

and

$$W^{out}_{(A,p1)} = O_{p1}/(O_{p1} + O_{p2}) = \frac{2}{5}$$

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

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \tag{7}$$

5. 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–JSpider–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. Evaluating the results: The algorithms are evaluated by comparing their results.

Normally, websites in different domains focus on different topics. Usually, the websites have rich linkages to describe the focused topics. On the other hand, they do a poor job describing non-focused topics. For example, the websites of most universities have a lot of information about scholarships and courses whereas the websites of travel companies mainly provide travel paths and scenes around

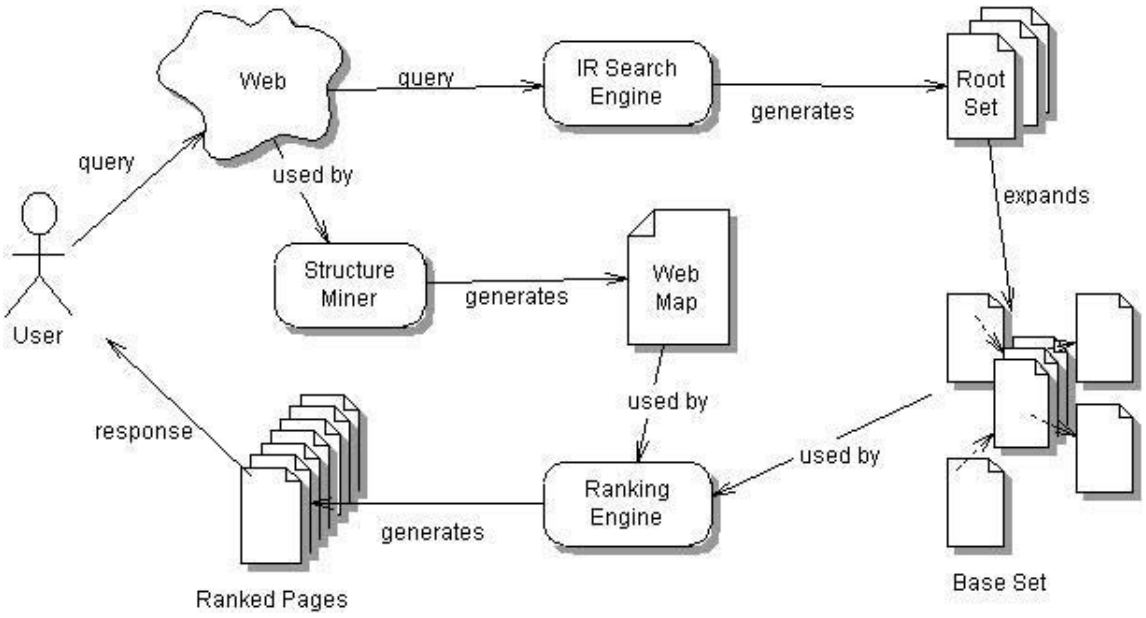

Figure 6. Architectural components of the system used to implement and evaluate the WPR algorithm

the world. To test the WPR algorithm for both focused and non-focused topics, we choose several queries from both categories. In this paper, an experiment using "travel agent," a non-focused topic in the website of Saint Thomas University, is provided in Appendices A and B.

6. Evaluation

The query topics "travel agent" and "scholarship" are used in the evaluation of the WPR and the standard PageRank 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 PageRank and the WPR algorithms provide important information about a given query by using the structure of the website. Some pages irrelevant to a given query are included in the results as well. For example, even though the home page of Saint Thomas University,
http://www.stu.ca/index.htm, is not related to the given query, it still receives the highest rank because of its many existing inlinks and outlinks. To reduce the noise resultant from irrelevant pages, 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.

An objective categorization of the results (lists of pages) is achieved by integrating the responses from several people: for each page, we compared the count of each category (i.e., VR, R, WR and IR) and chose the category with the largest count as the type of that page.

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

The performances of the WPR and the standard PageRank algorithms have been evaluated to identify the algorithm that produces better results (i.e., results that are more relevant to the given query). The WPR and the standard PageRank algorithms provide sorted lists (i.e.,

	Number of Relevant Pages		Relevancy Value(κ)	
Size of the 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

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

ranked pages) to users based on the given query. Therefore, in the result list, the number of relevant pages and their order are of great importance. The following rule has been adopted to calculate the relevancy value of each page in the list of pages.

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, κ, of a page-list is a function of its category and position:

$$\kappa = \sum_{i \in R(p)} (n - i) \times W_i \tag{8}$$

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 W_i is the weight of page i.

$$W_i = \begin{cases} \nu_1, & \text{if the ith page is VR} \\ \nu_2, & \text{if the ith page is R} \\ \nu_3, & \text{if the ith page is WR} \\ \nu_4, & \text{if the ith page is IR} \end{cases} \tag{9}$$

where $\nu_1 > \nu_2 > \nu_3 > \nu_4$.

The value of W_i for an experiment could be decided through experimental studies. For our experiment, we set ν_1, ν_2, ν_3 and ν_4 to 1.0, 0.5, 0.1 and 0, respectively, based on the relevancy of each category.

The relevancy values for the query "travel agent" are shown in Table 1. In this table, relevant pages represent the pages in the category VR as well as in the category R.

From Table 1, we see that WPR produces larger relevancy values, which indicate that WPR performs better than standard PageRank does. Figure 7 illustrates the performance. Moreover, the following two points are observed from Table 1:

- Within the first 10 pages, one relevant page is identified by WPR whereas no relevant page is determined by standard PageRank. This case indicates that WPR may be able to identify more relevant pages from the top of the result list than can standard PageRank.
- Within the first 20 pages, the relevancy value obtained from WPR is larger than that obtained from standard PageRank, even though one more relevant page is identified by standard PageRank. This scenario indicates that the relevant pages determined by WPR are either more relevant or ranked higher inside the list.

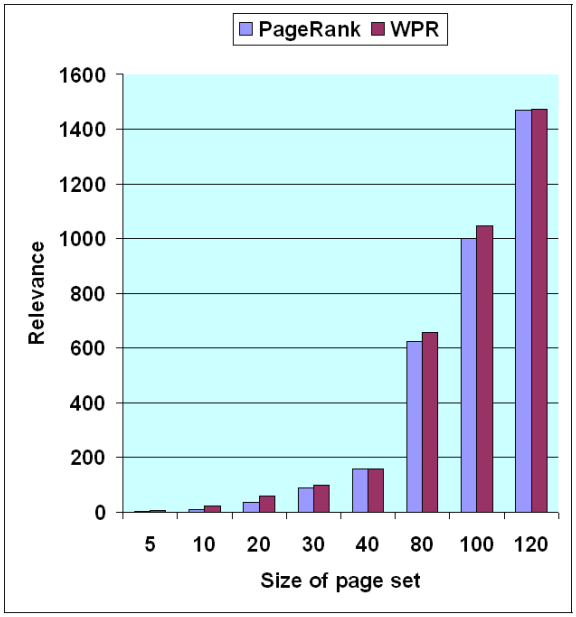

Figure 7. The relevancy value versus the size of the page set of the query "travel agent" for PageRank and WPR

	Number of Relevant Pages		Relevancy Value(κ)	
Size of the page set	PageRank	WPR	PageRank	WPR
5	2	3	2	5.5
10	2	4	9.5	22
20	4	4	34.5	57
30	8	5	87.5	99
40	10	8	158.5	159.3
80	16	15	624.8	655.3
100	22	19	999.2	1045.3
120	25	20	1470.4	1473.3

Table 2. The relevancy values for the query "scholarship" produced by PageRank and WPR using different page sets

6.3. Focused topic queries

This subsection evaluates the results obtained for the query "scholarship." This query is a focused topic within the website of Saint Thomas University. The relevancy values of the results are shown in Table 2.

Similar to the query "travel agent," Figure 8 demonstrates that the WPR algorithm produces better results (larger relevancy values) for the query "scholarship." Moreover, the two points derived from the query "travel agent" are shown more clearly in this case (see Table 2).

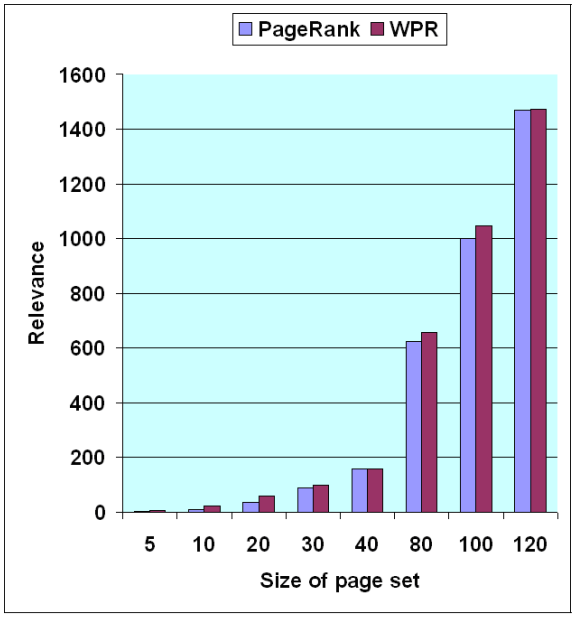

Figure 8. The relevancy value versus the size of the page set of the query "scholarship" for PageRank and WPR

In conclusion, the results obtained from WPR and standard PageRank for the focused and non-focused topics show that WPR is superior to standard PageRank.

7. 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. Simulation studies using the website of Saint Thomas University show that WPR is able to identify a larger number of relevant pages to a given query compared to standard PageRank.

In the current version of WPR, only the inlinks and outlinks of the pages in the reference page list are used in the calculation of the rank scores. In our future study of this method, we would like to consider the possibility of calculating the rank scores by using more than one level of reference page list. Moreover, a detailed analysis of WPR's performance using different websites and multiple levels of reference page lists would be carried out.

As part of our future work, we plan to carry out extensive performance analysis of WPR by using other web sites and increasing the number of 'human' users to categorize the web pages.

8. Acknowledgments

The authors graciously acknowledge the funding from the Atlantic Canada Opportunity Agency (ACOA) through the Atlantic Innovation Fund (AIF) and through grant RGPN 227441-00 from the National Science and Engineering Research Council of Canada (NSERC) to Dr. Ghorbani. The first author would also like to acknowledge the funding from the National Science and Engineering Research Council of Canada (NSERC). The authors would like to thank Mr. Elijah Bitting, Jie Zhang, and Lemin Wu for their help in categorizing the pages.

References

[1] 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, 1998.

[2] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, 1999.

[3] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proceedings of 17th International Conference on Machine Learning*, pages 167–174. Morgan Kaufmann, San Francisco, CA, 2000.

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

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

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

[7] 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.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Technical report,Stanford Digital Libraries SIDL-WP-1999-0120*, 1999.

[9] 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.

[10] C. Ridings and M. Shishigin. Pagerank uncovered. *Technical report*, 2002.

[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.

Appendix A: Result set of PageRank for ”travel agent”

Index	Category	URL of Page
1	IR	http://www.stu.ca/index.htm
2	IR	http://www.stu.ca/academic/scwk/rural/papers.htm
3	IR	http://www.stu.ca/academic/scwk/rural/programmeb.htm
4	IR	http://www.stu.ca/academic/scwk/rural/programme.htm
5	IR	http://www.stu.ca/academic/scwk/rural/presenters.htm
6	IR	http://www.stu.ca/academic/scwk/rural/list.htm
7	IR	http://www.stu.ca/academic/scwk/rural/eval.htm
8	IR	http://www.stu.ca/academic/scwk/rural/cheers.htm
9	WR	http://www.stu.ca/academic/scwk/rural/pugh.htm
10	IR	http://www.stu.ca/academic/scwk/rural/chart.htm
11	IR	http://www.stu.ca/academic/scwk/rural/registration.htm
12	WR	http://www.stu.ca/academic/scwk/rural/accommodation.htm
13	IR	http://www.stu.ca/academic/scwk/index.htm
14	VR	http://www.stu.ca/academic/scwk/rural/transport.htm
15	R	http://www.stu.ca/international/support.htm
16	WR	http://www.stu.ca/international/programmes.htm
17	WR	http://www.stu.ca/international/stu.htm
18	R	http://www.stu.ca/international/location.htm
19	VR	http://www.stu.ca/international/visas.htm
20	IR	http://www.stu.ca/international/finances.htm
21	IR	http://www.stu.ca/alumni/connections/fall2000/moore1.jpg
22	IR	http://www.stu.ca/international/images/button-say-default.gif
23	IR	http://www.stu.ca/international/images/button-location-default.gif
24	IR	http://www.stu.ca/international/images/button-finances-default.gif
25	IR	http://www.stu.ca/international/images/button-visas-default.gif
26	IR	http://www.stu.ca/international/images/button-stu-default.gif
27	IR	http://www.stu.ca/international/images/button-programmes-default.gif
28	IR	http://www.stu.ca/international/say.htm
29	IR	http://www.stu.ca/international/images/button-support-default.gif
30	WR	http://www.stu.ca/alumni/connections/fall2000/moore.htm
31	IR	http://www.stu.ca/international/
32	IR	http://www.stu.ca/international/images/logo-international.gif
33	IR	http://www.stu.ca/international/images/mainmenu-header.gif
34	IR	http://www.stu.ca/international/images/mainmenu-headerend.gif
35	IR	http://www.stu.ca/international/images/submenu-spacer.gif
36	IR	http://www.stu.ca/international/images/submenu-contacts.gif
37	IR	http://www.stu.ca/international/images/submenu-home.gif
38	IR	http://www.stu.ca/international/images/logo-copyright.gif
39	IR	http://www.stu.ca/admin/hr/polagree.htm
40	IR	http://www.stu.ca/admin/hr/policies/ptca.PDF
41	IR	http://www.stu.ca/international/default.css
42	IR	http://www.stu.ca/
43	IR	http://www.stu.ca/international/images/
44	IR	http://www.stu.ca/academic/scwk/rural/pics.htm
45	IR	http://www.stu.ca/international/images/arrow-location.gif
46	IR	http://www.stu.ca/international/images/button-location-over.gif
47	IR	http://www.stu.ca/international/images/header-location.jpg
48	IR	http://www.stu.ca/international/images/map-canada-off.gif
49	IR	http://www.stu.ca/international/images/map-nb-off.gif
50	IR	http://www.stu.ca/international/images/map-fredericton.gif
51	IR	http://www.stu.ca/international/images/submenu-location.gif
52	IR	http://www.stu.ca/international/index.htm
53	WR	http://www.stu.ca/inkshed/nlett302/ink19pgm.htm
54	VR	http://www.stu.ca/ hunt/k8/k8528.htm
55	WR	http://www.stu.ca/ hunt/33360102/crusoe.htm
56	IR	http://www.stu.ca/alumni/connections/fall2000/index.htm
57	WR	http://www.stu.ca/ rgmoore/fldofpep.htm
58	IR	http://www.stthomasu.ca/
59	IR	http://www.stu.ca/international/images/button
60	IR	http://www.stu.ca/academic/scwk/rural/index.htm
61	WR	http://people.stu.ca/ hunt/27730304/articles.htm
62	WR	http://people.stu.ca/ faulkner/crimstudyguide/USCoreSGChap12.htm
63	IR	file:/C—/index.htm
64	WR	http://people.stu.ca/ hunt/18c/33360102/finlwebs/gsqnv/satire.htm
65	WR	http://www.stu.ca/ hunt/27739900/articles.htm
66	IR	http://www.canada.gc.ca
67	WR	http://www.stu.ca/ hunt/27730102/articles.htm
68	IR	http://www.gov.nb.ca
69	VR	http://www.aircanada.ca/
70	VR	http://www.city.fredericton.nb.ca

Appendix B: Result set of WAPR for "travel agent"

Index	Category	URL of Page
1	IR	http://www.stu.ca/index.htm
2	IR	http://www.stu.ca/alumni/connections/fall2000/moore1.jpg
3	IR	http://www.stu.ca/academic/scwk/rural/papers.htm
4	IR	http://www.stu.ca/academic/scwk/rural/programmeb.htm
5	IR	http://www.stu.ca/academic/scwk/rural/programme.htm
6	IR	http://www.stu.ca/academic/scwk/rural/presenters.htm
7	IR	http://www.stu.ca/academic/scwk/rural/list.htm
8	IR	http://www.stu.ca/academic/scwk/rural/eval.htm
9	R	http://www.stu.ca/international/support.htm
10	WR	http://www.stu.ca/academic/scwk/rural/pugh.htm
11	IR	http://www.stu.ca/academic/scwk/rural/registration.htm
12	VR	http://www.stu.ca/academic/scwk/rural/transport.htm
13	IR	http://www.stu.ca/academic/scwk/rural/cheers.htm
14	IR	http://www.stu.ca/academic/scwk/rural/chart.htm
15	WR	http://www.stu.ca/academic/scwk/rural/accommodation.htm
16	WR	http://www.stu.ca/international/programmes.htm
17	WR	http://www.stu.ca/international/stu.htm
18	R	http://www.stu.ca/international/location.htm
19	WR	http://www.stu.ca/alumni/connections/fall2000/moore.htm
20	IR	http://www.stu.ca/international/
21	IR	http://www.stu.ca/international/images/
22	VR	http://www.stu.ca/international/visas.htm
23	IR	http://www.stu.ca/international/finances.htm
24	IR	http://www.stu.ca/admin/hr/polagree.htm
25	IR	http://www.stu.ca/academic/scwk/index.htm
26	IR	http://www.stu.ca/international/say.htm
27	IR	http://www.stu.ca/admin/hr/policies/ptca.PDF
28	IR	http://www.stu.ca/academic/scwk/rural/pics.htm
29	IR	http://www.stu.ca/international/images/button-location-default.gif
30	IR	http://www.stu.ca/international/images/button-support-default.gif
31	IR	http://www.stu.ca/international/images/button-stu-default.gif
32	IR	http://www.stu.ca/international/images/button-say-default.gif
33	IR	http://www.stu.ca/international/images/button-finances-default.gif
34	IR	http://www.stu.ca/international/images/button-visas-default.gif
35	IR	http://www.stu.ca/international/images/button-programmes-default.gif
36	IR	http://www.stu.ca/international/images/logo-international.gif
37	IR	http://www.stu.ca/international/images/mainmenu-header.gif
38	IR	http://www.stu.ca/international/images/mainmenu-headerend.gif
39	IR	http://www.stu.ca/international/images/submenu-spacer.gif
40	IR	http://www.stu.ca/international/images/submenu-contacts.gif
41	IR	http://www.stu.ca/international/images/submenu-home.gif
42	IR	http://www.stu.ca/international/images/logo-copyright.gif
43	IR	http://www.stu.ca/international/images/arrow-location.gif
44	IR	http://www.stu.ca/international/images/button-location-over.gif
45	IR	http://www.stu.ca/international/images/header-location.jpg
46	IR	http://www.stu.ca/international/images/map-canada-off.gif
47	IR	http://www.stu.ca/international/images/map-nb-off.gif
48	IR	http://www.stu.ca/international/images/map-fredericton.gif
49	IR	http://www.stu.ca/international/images/submenu-location.gif
50	IR	http://www.stu.ca/international/default.css
51	IR	http://www.stu.ca/
52	IR	http://www.stu.ca/international/index.htm
53	WR	http://www.stu.ca/inkshed/nlett302/ink19pgm.htm
54	VR	http://www.stu.ca/ hunt/k8/k8528.htm
55	WR	http://www.stu.ca/ hunt/33360102/crusoe.htm
56	IR	http://www.stu.ca/alumni/connections/fall2000/index.htm
57	WR	http://www.stu.ca/ rgmoore/fldofpep.htm
58	IR	http://www.stthomasu.ca/
59	IR	http://www.stu.ca/international/images/button
60	IR	http://www.stu.ca/academic/scwk/rural/index.htm
61	WR	http://people.stu.ca/ hunt/27730304/articles.htm
62	WR	http://people.stu.ca/ faulkner/crimstudyguide/USCoreSGChap12.htm
63	IR	file:/C—/index.htm
64	WR	http://people.stu.ca/ hunt/18c/33360102/finlwebs/gsqnv/satire.htm
65	WR	http://www.stu.ca/ hunt/27739900/articles.htm
66	IR	http://www.canada.gc.ca
67	WR	http://www.stu.ca/ hunt/27730102/articles.htm
68	IR	http://www.gov.nb.ca
69	VR	http://www.aircanada.ca/
70	VR	http://www.city.fredericton.ca