ISSN:0975-9646
Sandeep Gupta et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1478-1483

An Extended Algorithm of Page Ranking Considering Chronological Dimension of Search

Sandeep Gupta[#], Mohd. Husain[*]
[#]*Computer Science and Engineering,NIMS University, Jaipur, Rajasthan, India*
[*]*Director, AZAD IET, Lucknow, UP, India*

***Abstract*—The current generation of WWW search engines reportedly makes extensive use of evidence derived from the structure of the WWW to better match relevant documents and identify potentially authoritative pages. However, despite this reported use, there has been little analysis which supports the inclusion of web evidence in document ranking, or which examines precisely what its effect on search results might be. The success of document ranking in the current generation of WWW search engines is attributed to a number of web analysis techniques. 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. We introduce the Credence PageRank Algorithm (CPR) to improve the performance of web page ranking.**
As CPR is motivated by the observation that a hyperlink from a Web page to another is an implicit conveyance of authority to the target page, one can use these algorithms to find important Web pages. However, an important factor that is not considered by these techniques is the timeliness of search results. The Web is a dynamic environment. Quality pages in the past may not be quality pages now or in the future. Thus we also study search considering Chronological dimension.

***Keywords*—Web Mining, Web Usage Mining, Page Rank, Web Map**

I. INTRODUCTION

During the past few years the World Wide Web has turn into the biggest and most admired way of communication and information dissemination. It serves as a platform for exchanging various breeds of information, ranging from research papers, and educational content, to multimedia content, software and personal logs (blogs). Every day, the web grows by roughly a million electronic pages, adding to the hundreds of millions pages already on-line. Because of its rapid and chaotic growth, the resulting network of information lacks of organization and structure. Users often feel disoriented and get lost in that information overload that continues to expand. On the other hand, the e-business sector is rapidly evolving and the need for web market places that anticipate the needs of their customers is more than ever evident. Therefore, the ultimate need nowadays is that of predicting the user needs in order to improve the usability and user retention of a web site.

II. HYPERLINK RECOMMENDATION

A. *Link counting / in-degree*

A page's in-degree score is a measure of its direct prestige, and is obtained through a count of its incoming links. It is widely believed that a web page's in-degree may give some indication of its importance or popularity [2].

B. *PageRank*

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[3][[4]. Figure 1 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.

A slightly simplified version of PageRank [3][4] is defined as

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

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. 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. To solve the rank sink problem, we observed the users' activities. A phenomenon is found that not all users follow the existing links[7]. 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.

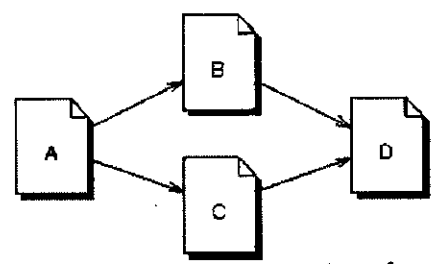

Figure 1. An example of backlinks

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.

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

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 [4]. 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)[3][4].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 exists: in the actual web, some links in a web page may be more important than are the others.

III. CREDENCE PAGERANK (CPR)

The more popular web pages are, the more linkages that other webpages tend to have to them or are linked to by them. The proposed extended 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 $C^{in}_{(v,u)}$ and $C^{out}_{(v,u)}$ respectively.

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

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

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.

$C^{out}_{(v,u)}$ is the credit 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.

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

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 2 shows an example of some links of a hypothetical website.

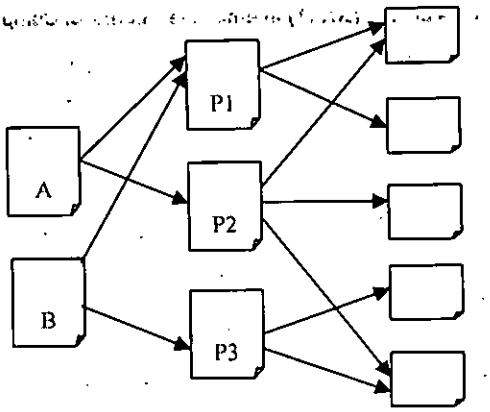

Figure 2. Links of a website

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,

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

and

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

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

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

IV. EXPERIMENTAL EVALUATION

A. Methodology

To evaluate the CPR algorithm, we implemented CPR and the standard PageRank algorithms to compare their results. Figure 6 illustrates different components involved in the implementation and evaluation of the CPR 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 CPR algorithms rely on the web structure. After comparing the structures of several web sites of various colleges and universities, the website of Azad Institute of Pharmacy and Research(AIPR), Lucknow (www.aipr.ac.in) from Azad Group of Institutions, Lucknow and Teerthankar Mahaveer University (www.tmu.ac.in) has been chosen.

2) Building a web map

Using JSpider (An open source spider software)[5] and WebTracer (A software tool to visualize the structure of the web)[6] is used to generate the required web map. The link structure and web map of Azad Institute of Pharmacy and Research (AIPR) website and Web map of Teerthankar Mahaveer University website are shown in figure 3 and 4.

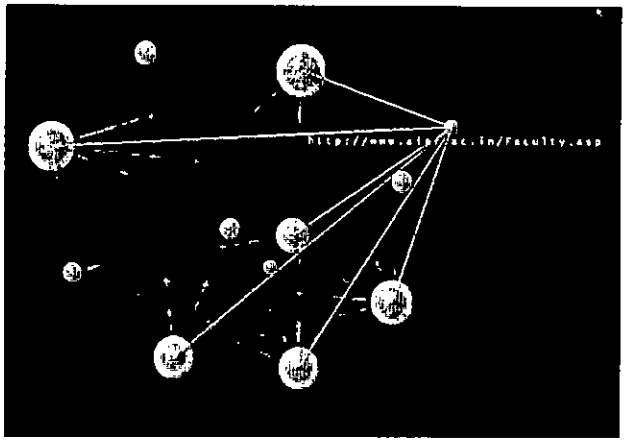

Figure 3 (a) Web Map of aipr.ac.in

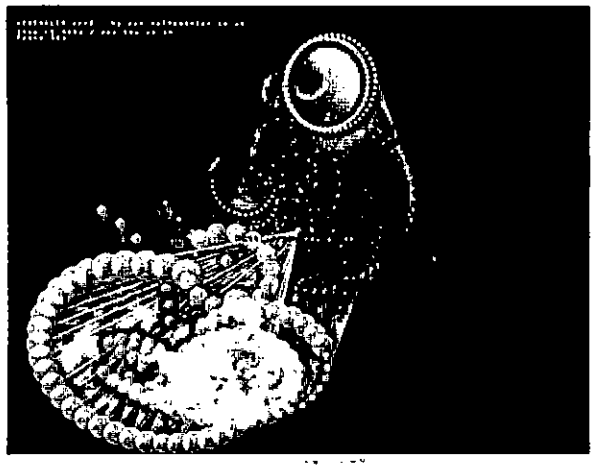

Figure 3(b) Web map of www.tmu.ac.in

```
musthandle="true">
- <folders>
  + <folder name="style">
  - <folder name="Scripts">
    - <resources>
      - <resource file="AC_RunActiveContent.js"
          url="http://www.aipr.ac.in/Scripts/AC_RunActiveContent.js">
          <state>PARSE_IGNORED</state>
        - <info>
            <http-status>200</http-status>
            <size>8321</size>
            <mime>application/x-javascript</mime>
            <fetch-time>8406</fetch-time>
          </info>
        - <referers>
            <referer url="http://www.aipr.ac.in/index.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in" count="1" />
            <referer url="http://www.aipr.ac.in/Infra.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/Faculty.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/PhotoGallery.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/Courses.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/AboutUs.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/Admission.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/Placement.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/Events.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/ContactUs.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/DirectorMessage.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/vision.asp"
              count="1" />
            <referer url="http://www.aipr.ac.in/link.htm" count="1" />
            <referer url="http://www.aipr.ac.in/Fee.asp" count="1" />
          </referers>
        </resource>
      </resources>
      <folders />
    </folder>
  - <folder name="aipr">
    + <resources>
```

Figure 4. Link Structure output of aipr.ac.in by using Jspider

3). Finding the root set

A search engine was developed for finding whether a webpage is relevant, very relevant, week relevant or irrelevant corresponding to a keyword (query). This search engine was developed in .Net Environment using C#. We named it as Relevancy Search Engine. The screenshot of output is shown in figure 5.

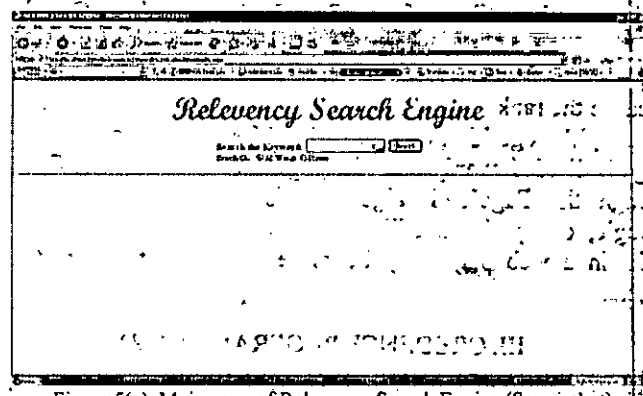

Figure 5(a). Main page of Relevancy Search Engine (Screenshot)

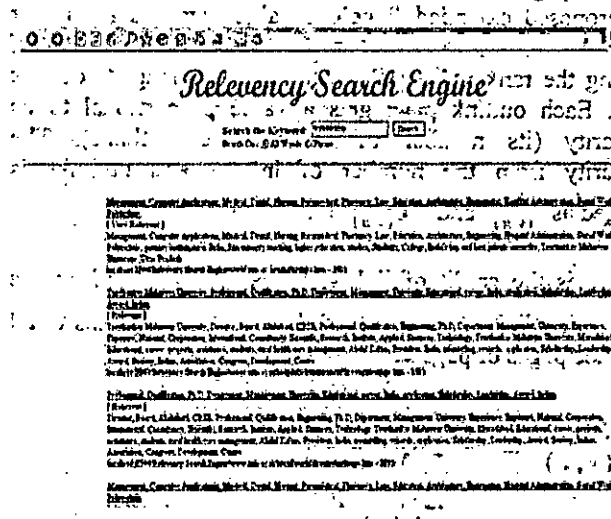

Figure 5(b). Search Result

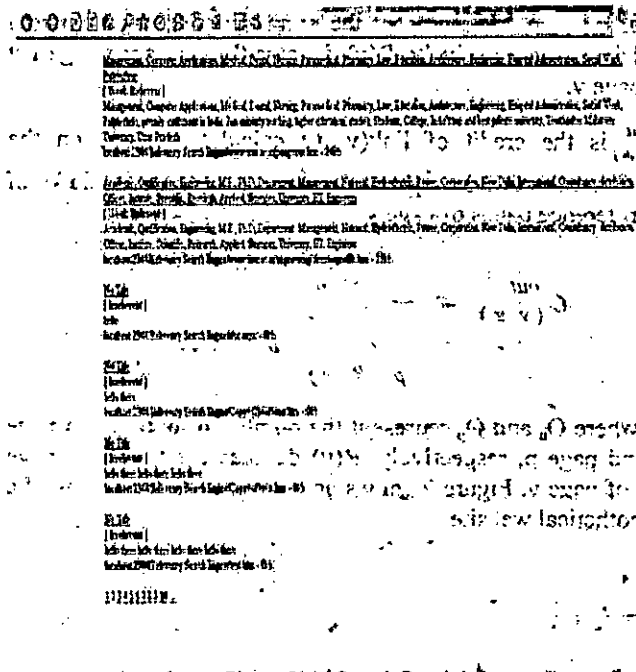

Figure 5(c). Search Result

A set of pages relevant to a given query is retrieved using this search engine. We aim to embed this search engine in the web site. We call rootset to this set of pages

4) Finding the base set

By the help of web map we create a base set 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 CPR 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 the world. To test the CPR algorithm for both focused and non-focused topics, we choose several queries from both categories.

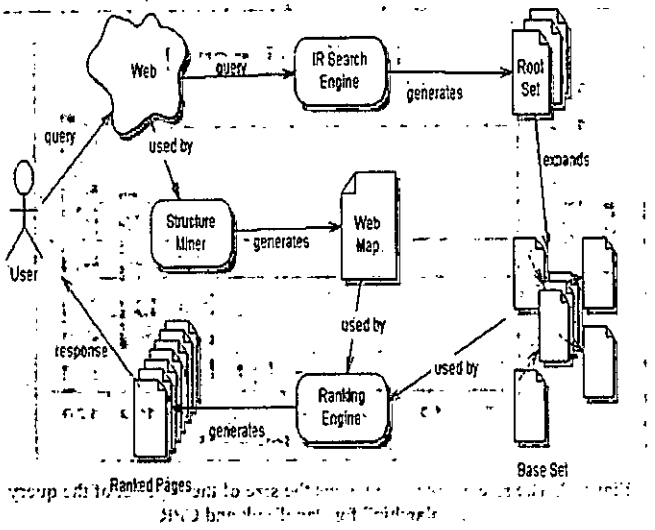

Figure 6. Architectural components of the system used to implement and evaluate the CPR algorithm

B. Evaluation

The query topics "travel agent" and "scholarship" are used in the evaluation of the CPR and the standard PageRank algorithms. "Travel agent" represents a non-focused topic whereas "scholarship" represents a focused (popular) topic in the website of TMU, Moradabad. The results of the evaluation are summarized in the following subsections.

1). The determination of the relevancy of the pages to the given query

The Standard PageRank and the CPR 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 Teerthankar Mahaveer University (TMU), Moradabad (www.tmu.ac.in) 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.

2). The Calculation of the relevancy of the page lists to the given query

The performances of the CPR 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 CPR and the standard PageRank algorithms provide sorted lists (i.e., 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.

3). 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, K, of a page-list is a function of its category and position:

$$K = \sum_{i \in R(p)} (n - i) \times C_i$$

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 Ci is the weight of page i.

$$C_i = \begin{cases} V_1, if\ the\ i^{th}\ page\ is\ VR \\ V_2, if\ the\ i^{th}\ page\ is\ R \\ V_3, if\ the\ i^{th}\ page\ is\ WR \\ V_{41}, if\ the\ i^{th}\ page\ is\ IR \end{cases}$$

where v1 > v2 > v3 > v4.

The value of C_i for an experiment could be decided through experimental studies. For our experiment, we set v_1, v_2, v_3, and v4 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 CPR produces larger relevancy values, which indicate that CPR performs better than standard PageRank does. Figure 7 illustrates the performance. Moreover, the following two points are observed from Table 1:

TABLE 1

THE RELEVANCY VALUES FOR THE QUERY "TRAVEL AGENT" PRODUCED BY PAGERANK AND CPR USING DIFFERENT PAGE SETS

Size of the page set	Number of Relevant Pages		Relevancy Value(κ)	
	Page Rank	CPR	Page Rank	CPR
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

- Within the first 10 pages, one relevant page is identified by CPR whereas no relevant page is determined by standard PageRank. This case indicates that CPR 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 CPR 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 CPR are either more relevant or ranked higher inside the list.

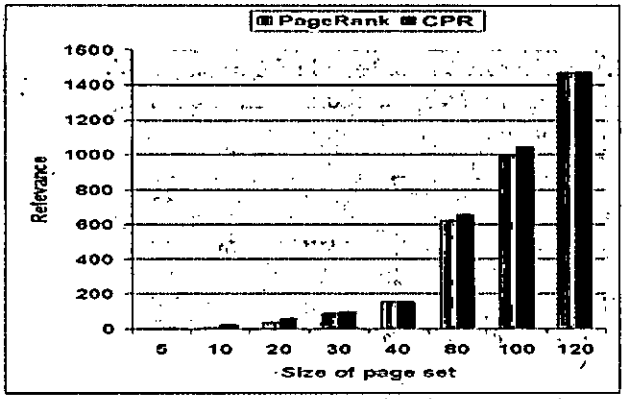

Figure 7. The relevancy value versus the size of the page set of the query "travel agent" for PageRank and CPR

4). Focused topic queries

This subsection evaluates the results obtained for the query "scholarship." This query is a focused topic within the website of Teerthankar Mahaveer University (TMU), Mooradabad (www.tmu.ac.in). The relevancy values of the results are shown in Table 2.Similar to the query "travel agent," Figure 8 demonstrates that the CPR 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).

TABLE 2

THE RELEVANCY VALUES FOR THE QUERY "SCHOLARSHIP" PRODUCED BY PAGERANK AND CPR USING DIFFERENT PAGE SETS

Size of the page set	Number of Relevant Pages		Relevancy Value(κ)	
	Page Rank	CPR	Page Rank	CPR
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

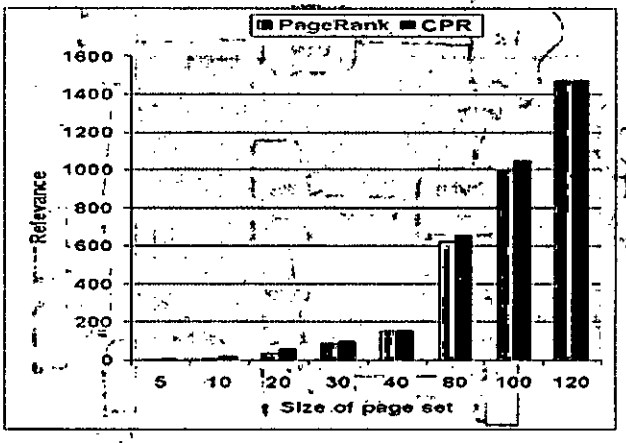

Figure 8. The relevancy value versus the size of the page set of the query "scholarship" for PageRank and CPR

V. EXTENSION OF CPR CONSIDERING CHRONOLOGICAL DIMENSION

We believe that dealing with the problems related to the Chronological dimension of search is of great importance to future developments of search technology. In our research, we take a step towards this direction. To understand the issues in greater detail, we coarsely classify Web pages into two types; old pages and new pages.

- Old pages: These are the pages that have appeared on the Web for a long time. We can also classify these pages into quality pages and common pages.
- Quality pages usually have a large number of in-links, and common pages do not have many in-links. Old quality pages can be further classified from the temporal dimension:
- Old quality pages that are up-to-date: As time goes by, the authors of the pages update the contents to reflect the latest developments. Such pages often stay as quality pages.

- Old quality pages that are not up-to-date: The authors of these pages do not update their contents over time. These pages become outdated, and receive fewer and fewer new in-links over time. However, if many Web users do not clean up hyperlinks to these pages, they may still maintain a sizeable in-links, and would be ranked high in spite of their low value. Regarding old common pages, we can also classify them into two types from the chronological dimension:
  - Old common pages that remain common pages: Most pages on the Web are such pages. As time goes by, they are still common pages, as they do not receive many in-links.
  - Old common pages that have become important: These pages were not important in the past, but as time goes by they become valuable pages. This transition may be due to a number of reasons, such as fashion change or quality contents being added.
- New pages: These are pages that appeared on the Web recently. New pages can also be grouped into categories:
  - New quality pages: These pages are new and are of high quality. However, they received few or no inlinks because they are new.
  - New common pages: These pages are new and common. Unlike an older page, a new page receives few or no in-links. It is thus difficult to judge if it is a quality page.

In context with research papers timing factors are as follows:

1. The publication date, and
2. The dates that the paper is cited by other papers.

We now describe the Chronological PageRank technique. Since we are interested in the importance of a paper now, a citation occurred a few months ago is clearly more important than one occurred a few years ago. We modify the PageRank technique by crediting each citation. The system calculates the Chronological PageRank (PR^T) value for each paper as follows:

$$PR^T(A) = (1-d) + d \times \left(\frac{C_1 \times PR^T(p_1)}{C(p_1)} (2) \cdots + \frac{C_n \times PR^T(p_n)}{C(p_n)} \right)$$

In this equation, c_i is the time based credit for each citation. Its value depends on the citation date from paper p_i to A, which is also the publication date of p_i . The earlier the citation occurred, the smaller the credit is. Since exponential average is extensively used in time-series prediction, we choose to decay the credits exponentially according to time,

$$c_i = DecayRate^{(y-t_i)/12}$$

where y is the current time, t_i is the publication time of paper p_i ; and $(y-t_i)$ is the time gap in months. *DecayRate* is a parameter.

VI. 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 CPR algorithm, an extension to the PageRank algorithm. CPR 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 AIPR, Lucknow and website of TMU, Mooradabad show that CPR is able to identify a larger number of relevant pages to a given query compared to standard PageRank.

We also studied the chronological dimension of search. So far, limited research work has been done to consider time in either publication search or Web search.

REFERENCES

[1] Raj Gaurang Tiwari et al. "A Quantitative Approach to Perk up Navigation Efficiency" in proceedings of 1st International Conference On Management Of Technologies & Information security (ICMIS 2010), held at IIIT, Allahabad, India, Jan 21st - 24th, 2010, ISBN No. 978-81-8329-375-4, pp-378-386.

[2] ZHU , X., AND GAUCH , S. Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web. Tech. rep., Department of Electrical Engineering and Computer Science, University of Kansas,2000.

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

[4] Ridings and M. Shishigin. Pagerank uncovered. Technical report, 2002.

[5] Jspider http://www.j-spider.sourceforge.net/

[6] Webstracer2, www.nullpointer.co.uk/-/webtracer2.htm

[7] Raj Gaurang Tiwari et al. "Recuperating Website Link Structure Using Fuzzy Relations between the Content and Web Pages", in International Journal International Journal of Computer Applications, Vol. 1, 2010, URL http://www.ijcaonline.org/archives/number1/245-402