A Comparative Study of Page Ranking Algorithms for Online Digital Libraries

Sumita Gupta, Neelam Duhan, Poonam Bansal

Abstract— With the increasing use of academic digital libraries, it becomes more important for authors to have their publications or scientific literature well ranked in order to reach their audience. Web mining is a potential candidate to meet this challenge. These techniques when applied on the search results of a user query, provide them an order so that users can easily navigate through the search results and find the desired information content. This is also called Page Ranking. The ranking mechanisms can arrange the scientific literature or publications in order of their relevance, importance and content score. In this paper, a survey of some prevalent page ranking algorithms for academic digital libraries is being carried out and comparison of these algorithms in context of performance has been done.

Keywords: Digital libraries, Search engine, Page ranking, Web mining

1 Introduction

With the rapid growth of information sources available on the World Wide Web (WWW) and growing needs of users, it is becoming difficult to manage the information on the Web and satisfy the user needs [1]. For this purpose, many advanced web searching and mining techniques have recently been developed to tackle the problem of finding relevant information and are being used in the commercial search engines such as Google and Yahoo.

In spite of advances in search engine technologies, there still occur situations where the user is presented with non-relevant search results. For example, when a user inputs a query for some scientific literature, book or periodical to a general purpose search engine such as Google, it returns a long list of search results consisting of tutorials, news, articles, blogs etc. This happens due to limited crawling by the search engines. Most of the search engines are not completely capturing the vast amount of information available in the digitization projects on books and periodicals that are occurring locally, nationally and internationally.

Moreover, researchers are making their work available online in the form of postscript or PDF documents, therefore, amount of scientific information and the number of electronic journals on the Web is increasing at a fast rate. But the access to the growing body of scientific literature on the publicly indexable Web is limited by the lack of organization of such information. To overcome this problem, digital libraries have been introduced to make retrieval mechanism more effective and relevant for researchers or users. A digital library [2] is an integrated set of services for capturing, cataloging, storing, searching, protecting and retrieving information, which provides coherent organization and convenient access to typically large amounts of digital information. Now a day, digital libraries are experiencing rapid growth with respect to both the amount and richness of available digital content. As a consequence of the availability of huge amounts of digital content, modern search engine technologies are now being introduced in digital libraries to retrieve the relevant content.

The architecture of a typical digital library search system is shown in Fig.1. The main component of this system is a crawler that traverses the hypertext structure in the web, downloads the web pages or harvest the desired papers published in specific venue (e.g. a conference or a journal) and stores them in database. Usually the publications present on WWW are in the form of postscript files or PDF. Therefore, for every user search, a new instance of the agent is created which locates and downloads postscript files identified by ".ps", ".ps.Z", or ".ps.gz" extensions. These downloaded files are passed through the document parsing sub agent which extracts the semantic features from the downloaded documents and places them into a database as parsed documents. The parsed documents are routed to an indexing module that builds the index based on the keywords present in the pages.

When the user fires a query in the form of keywords on the search interface of a digital library search system, it is retrieved by the database search and browsing sub agent which does query processing by taking the user query in proper syntax and returns an HTML formatted response to the user. The search results are usually presented in the form of an ordered list by the application of page ranking algorithms employed by digital libraries.

In this paper, a survey of some prevalent page ranking algorithms for online academic digital libraries has been done and a comparison is carried out. This paper is structured as follows: in Section II, web mining concepts, categories and technologies have been discussed. Section III provides a detailed overview of some page ranking algorithms with their strengths and weaknesses. Section IV presents an extensive comparison study. Finally in Section V, conclusion is drawn with a light on future suggestions.

2 Web MininG

Web mining [3, 4] is a means for automatically discovering and explore useful information from the WWW. There are three areas of Web Mining according to the web data used as input. These are Web Content

- *Sumita Gupta is currently pursuing Ph.D in Computer Engineering from YMCA University of Science & Technology, Faridabad, India, E-mail: sumitagoyal@gmail.com*
- *Neelam Duhan, Associate Professor, YMCA University Science & Technology, Faridabad, India*
- *PoonamBansal, Associate Professor, Maharaja Surajmal Institute of Technology, Delhi, India*

Mining (WCM), Web Usage Mining (WUM), and Web Structure Mining (WSM).

Web Content Mining (WCM) is a process of scanning and mining the text, pictures and graphs of a Web page to determine the relevance of the content of the web page in accordance to the search query.

Web Structure Mining (WSM) tries to discover the useful knowledge from the structure of hyperlinks catalogs them and generates information such as the similarity and relationship between papers by taking advantage of their hyperlink topology. WSM uses the graph theory to analyze node and connection structure of a web site where web pages act as nodes and hyperlinks as edges connecting two related pages.

The goal of WSM is to generate structured summary about the website and web page. Table 1 gives an overview of the three mining categories [5].

Web Usage Mining (WUM) is a process of identifying the browsing patterns by analyzing the user's navigational behavior while surfing on the Web. It extracts data stored in server access logs, referrer logs, agent logs, client-side cookies, user profile and Meta data.

3 PAGE RANKING

Today, the main challenge in front of search engines is to efficiently harness scientific work present on the WWW and present relevant results to the user. Web mining techniques are used in order to extract the relevant information and order the documents.

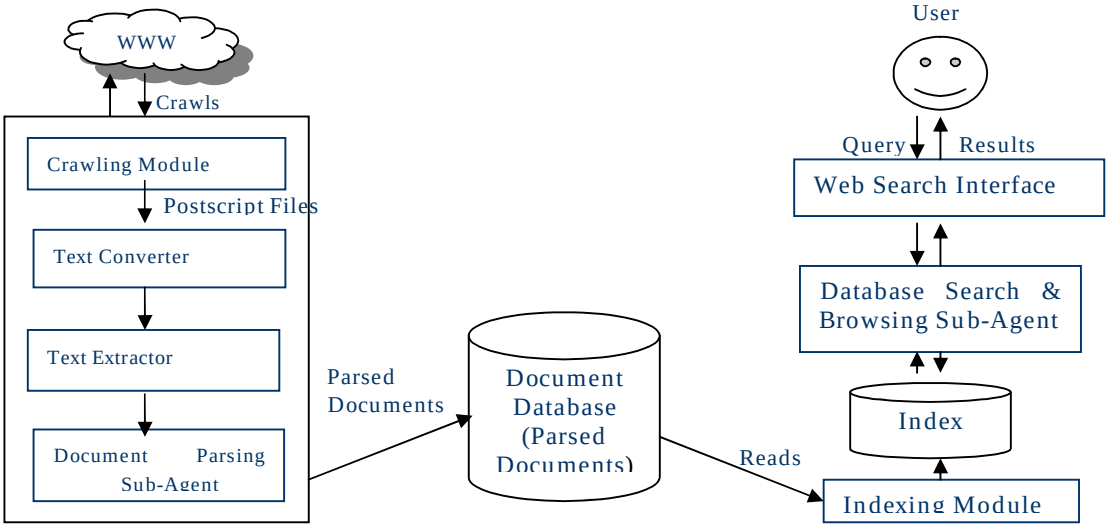


Fig.1. The Architecture of a Digital Library Search System

Table 1. Web Mining Categories

	Web Mining			
	Web Content Mining(WCM)		Web Structure Mining (WSM)	Web Usage Mining (WUM)
	Information Retrival View	DataBase View		
View of Data	-Unstructured -Structured	-Semi Structured -Web Site as DB	-Link Structure	-Interactivity
Main Data	- Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Server Logs -Browser Logs
Representation	-Bag of words, n-gram Terms, -phrases, Concepts or ontology -Relational	-Edge labeled Graph, -Relational	-Graph	-Relational Table -Graph
Method	-Machine Learning -Statistical (including NLP)	-Proprietary algorithms -Association rules	-Proprietary algorithms	-Machine Learning -Statistical -Association rules
Application Categories	-Categorization -Clustering -Finding extract rules -Finding patterns in text -User Modeling	-Finding frequent sub structures -Web site schema discovery	-Categorization -Clustering	-Site Construction -adaptation and management -Marketing -User Modeling

To represent the documents in an ordered manner, Page ranking methods are applied which can arrange the documents in order of their relevance and importance. Some of the common page ranking algorithms for online digital libraries have been discussed here as follows.

2.1 Citation Count Algorithm

This is one of the most frequent used ranking algorithm for measuring a scientist's reputation, and named as Citation Count [6]. This method uses the citation graph of the web to determine the ranking of scientific work. In citation graph, the nodes represent publications, whereas an edge from node i to node j represent a citation from paper i to paper j i.e. a vote from paper i to paper j. This method states that if a publication has more number of citations (incoming links) to it then publication become important. Therefore, it takes backlinks into account to order the publications. Thus, a publication obtains a high rank if the number of its backlinks is high. Citation Count is defined in (1):

$$CC_i = |I_i| \quad (1)$$

where CC_i represents the citation count of publication i, $|I_i|$ denotes the number of citations (in-degree) of the publication i.

Example Illustrating Working of CC. To explain the working of Citation Count, let us take an example of citation graph as shown in Fig. 2 and Table 2, where A, B , C, D, E and F are six publications.

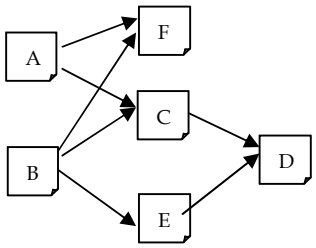


Fig. 2 Example of Citation Graph

Table 2. Data of citation Graph

Publication	Publication year
A	2011
B	2008
C	1998
D	1980
E	2007
F	2000

The Citation Count for publications A, B, C, D, E and F can be calculated by using (1):
CC(A)=0, CC(B)=0, CC(C)=3, CC(D)=2, CC(E)=1, CC(F)=2

The ranking of publications based on Citation Count become:
CC (C) > (CC (D), CC (F)) > CC (E) > (CC (A), CC (B))

Limitations of CC. There are a number of cases where this method fails to reveal the good picture of influence of publications in its domain. Few of reasons for this are:

- It does not take into account the importance of citing paper i.e. citation from the reputed journal get the equal weightage as the citation from the poor journal.
- If two papers have similar citation count as publication D and publication F shown in Fig 2, but interestingly publication F is almost 20 years younger than the publication D, thus it had a much smaller time window to accumulate citations. Thus, it does not take into consideration different characteristics of the citations, like their publication date.

2.2 Time dependent Citation Count Algorithm

Ludmila Marian [7, 8] proposed an extension to standard Citation Count method named Time Dependent Citation Count (TDCC). It is a time-dependent approach which takes into account time of the citation. This method assumes that the freshness of citations and link structure are factors that need to be taken into account in citation analysis while computing the importance of a publication. Thus, Citation Count algorithm is modified by initially distributing random surfers exponentially with age, in favor of more recent publications. The method introduces the effect of time in the citation graph by applying a time-decay factor to the citation counts. The weight of a publication i is denoted as $Weight_i$as given in (2)

$$Weight_i = e^{-w(t_p - t_i)} \quad (2)$$

where t_i denotes the published year of publication i, t_p denotes the present time (i.e. year), and w denotes the time decay parameter (w ϵ (0, 1]), which quantifies the notions of "new" and "old" citations (i.e. publications with ages less than the time decay parameter would be considered "new"; publications with ages larger than the time decay parameter would be considered "old") citations (in-degree) of the publication i.

Example Illustrating Working of TDCC. To illustrate the working of TDCC, let us refer again to Fig 2 and Table 1. By using (2) weight scores of publications can be calculated as:

$$Wt_A = 0 \quad (2a)$$

$$Wt_B = 0 \quad (2b)$$

$$Wt_C = e^{-w(2012-2011)} + e^{-w(2012-2008)} + e^{-w(2012-2000)} = e^{-w(1)} + e^{-w(4)} + e^{-w(12)} \quad (2c)$$

$$Wt_D = e^{-w(2012-1998)} + e^{-w(2012-2007)} = e^{-w(14)} + e^{-w(5)} \quad (2d)$$

$$Wt_E = e^{-w(2012-2008)} = e^{-w(4)} \quad (2e)$$

$$Wt_F = e^{-w(2012-2011)} + e^{-w(2012-2008)} = e^{-w(1)} + e^{-w(4)} \quad (2f)$$

where w is time decay factor. Let us take the threshold age = 6 years i.e. w=0 for the publications with the ages less than 6 years (considered new publications) and w=1 for publications with ages more than 6 years (considered old publications). By calculating the above equations, the rank score of publications become:

TDCC (A) = 0, TDCC (B) = 0, TDCC (C) =2.0000006144
TDCC (D) = 1.000000832, TDCC (E) = 1, TDCC (F) = 2

Here TDCC(C)>TDCC (F) > TDCC (D) > TDCC (E) > (TDCC (A), TDCC (B)). It may be noted that the resulting ranking of citations obtained by CC and TDCC is different.

Advantages and Limitations of TDCC. After adding a time decay parameter, the time-dependent ranking can differentiate between an old publication that acquired a large number of citations over a long period of time, and a new publication. That, although important for the scientific community, did not have enough time to acquire as many citations as the old one, in the favor of the latter.

- Adding a week or strong time decay factor to a ranking method will have an impact on the final ordering of the documents. For example adding a strong time decay factor to ranking will reveal the most popular publications at the current moment in time.
- Like CC, this method does not take into consideration the different importance of each citation have been discussed here as follows.

2.3 PageRank Algorithm

Surgey Brin and Larry Page [9,10] proposed a ranking algorithm, named PageRank (PR) which extends the idea of citation analysis. In citation analysis, the incoming links are treated as citations which provide importance to a page but this technique could not provide fruitful results. In turn, PageRank [10] provides a better approach which is based on the fact, that the importance of a research paper can be judged by the number of citations the paper has from other research papers. This algorithm states that if a link comes from an important paper then this link is given higher weightage than those which are coming from non-important papers. These links are called as backlinks. The PageRank of a paper u can be calculated as:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \qquad (3)$$

where u represents a paper, B(u) is the set of papers that point to u, PR (u) and PR (v) are rank scores of papers u and v respectively, N_v denotes the number of outgoing links of paper v, and d is a normalization factor.

Example Illustrating Working of PR. Let us take a previous example as shown in Fig 2 in order to explain the working of PageRank algorithm. The PageRanks for papers can be calculated by using (3):

$$PR(A) = (1-d) + d(0) \qquad (3a)$$
$$PR(B) = (1-d) + d(0) \qquad (3b)$$
$$PR(C) = (1-d) + d\left(\frac{PR(A)}{2} + \frac{PR(B)}{3} + \frac{PR(F)}{1}\right) \qquad (3c)$$
$$PR(D) = (1-d) + d\left(\frac{PR(C)}{1} + \frac{PR(E)}{1}\right) \qquad (3d)$$
$$PR(E) = (1-d) + d\left(\frac{PR(B)}{3}\right) \qquad (3e)$$
$$PR(F) = (1-d) + d\left(\frac{PR(A)}{2} + \frac{PR(B)}{3}\right) \qquad (3f)$$

Table 3 Iteration Method for PageRank

Iterations	PR (A)	PR (B)	PR (C)	PR (D)	PR (E)	PR (F)
0	1	1	1	1	1	1
1	0.15	0.15	1.106	1.090	0.192	0.256
2	0.15	0.15	0.474	0.552	0.192	0.256
3	0.15	0.15	0.474	0.552	0.192	0.256

Let us assume the initial PageRank as 1, d is set to 0.85 and do the calculation. The rank values of papers are iteratively substituted in above page rank equations to find the final values until the page ranks get converged as shown in Table 3.

By calculating the above equations iteratively, the page ranks of papers become:

PR (D) > PR (C) >PR (F) > PR (E) > (PR (A), PR (B))

Advantages and Limitations of PR. One of the main advantages of this method is that it ranks the publications accordingly to the importance of their citations, bringing to light some very insightful publications that would not have been discovered with the Citation Count method. On the other hand, there are some shortcomings of this ranking method also [11]:

- The rank score of publication is equally distributed among its all references irrespective of assigning the larger rank values to more important papers.
- A page rank of a publication is mostly affected by the scores of the publications that point to it and less by the number of citations. For example, in Fig. 3, node F gets higher score than node E, although node E gets 4 citations and node F gets 1 citation.
- PageRank gives high score to a node u, if it contained a cycle. For Example, Table 4 shows the rank results of graph shown in fig 3. In this, node E gets 4 citations, whereas node T gets 3 citations. However, the PageRank score of node T is about 2 times higher than that of node E. This happens because node T is a part of citation cycle. But in bibliometrics, cycles represents the self-citations which do not occur in citation graph. Thus PageRank does not provide fruitful results in bibliometrics.

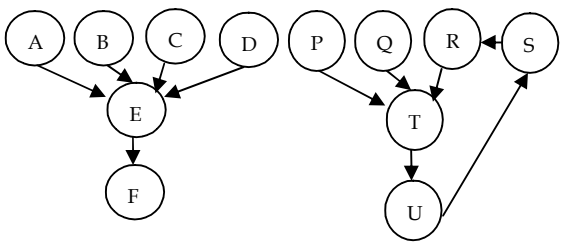


Fig. 3. An example of a graph

Table 4 Rank results Fig. 3

Node	CC	PR
A	0	0.15
B	0	0.15
C	0	0.15
D	0	0.15
E	4	0.66
F	1	0.71
P	0	0.15
Q	0	0.15
R	1	1.15
S	1	1.28
T	3	1.38
U	1	1.32

2.4 Popularity Weighted Ranking Algorithm

Yang Sun and C. Lee Giles [12] gave a new ranking method based on PageRank with significant improvement for ranking academic papers, named Popularity Weighted Ranking algorithm. This method combines the concepts that seem to be important for analyzing the importance of publication. The publication importance is determined on the basis of the weighted citations from the other papers and a popularity factor of its publication venue i.e. quality of the publication venue where a publication is published. Unlike impact factor, it does not differentiate between journals, conferences and workshop proceedings. The popularity factor of a publication venue v in a given year is defined by (4)

$$PF\ (v,t)\ =\ \frac{\mathrm{n_v}}{\mathrm{N}} \times \sum_{\mathrm{i}\in \mathrm{P}} \frac{\mathrm{PF\ (i,t)\ \times w(t)}}{\mathrm{N(t)}} \qquad (4)$$

where PF (v,t) represents the popularity factor of publication venue v in a given year t, P represents the set of publication venues i which cite v in that year, n_v denotes the number of papers published in venue v in that year, w (i) is the weight which represents the frequency that venue i cites venue v and N (i) denotes the total number of references generated by venue i. Considering the importance of popularity factor of publication venue, the ranking score of publication p at a previous time t is given in (5).

$$R\ (q_T)\ =\ \mathrm{PF}(\mathrm{v_{PT}})\ +\ \sum_{\mathrm{t}>T\ ,\mathrm{q_t}\in \mathrm{D}} \frac{\mathrm{R(q_t)}}{\mathrm{N(q_t)}} \qquad (5)$$

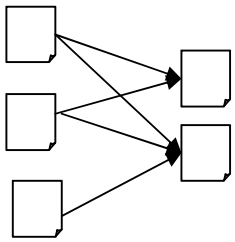
where R(q_t) represents the ranking score of a paper q_t, which is published at time t and cite paper p_T, D represents the set of papers which cite p_T, N(q_t) denotes the number of references in paper q_t, PF(v_{p_T}) denotes the popularity factor of the publication venue v where paper p_T is published.

Advantages and Limitations of Popularity Weighted Ranking Algorithm. One of the main advantages of this method is that it overcomes the limitations of impact factor i.e. by considering the impact of all publication venues and the probability of reader access.

- This algorithm works well for most queries but it does not work well for others.
- This method assumes that ranking score of a previously published paper will not have any impact on later published ones i.e. it does not take into consideration the time of publication.
- This method also does not differentiate between the popular and prestigious author who published the papers.

2.5 HITS Algorithm

Kleinberg [13] proposed a more refined notion for the importance of the web pages called Hyperlink Induced Topic Search (HITS). This method identifies two different forms of Web pages called hubs and authorities. Authorities are pages having important contents and hubs are pages that act as resource lists, guiding users to authorities as shown in Fig 1. A good authority is a page pointed to by good hubs, while a good hub is a page that points to good authorities. A page may be a good hub and a good authority at the same time.



F
i

Fig 4: Hubs and Authorities

Input: Root set R;
Output: Base set S
Let S = R
1.For each page p ϵ S, do Steps 3 to 5
2. Let T be the set of all pages S points to.
3. Let F be the set of all pages that point to S.
4. Let S = S + T + some or all of F.
5. Delete all links with the same domain name.
6. Return S

Fig 5: Algorithm to determine Base Set

HITS functions in two major steps.

1. Sampling Step: Sampling Step: In this step, a set of relevant pages for the given query are collected i.e. a sub-graph S of G is retrieved which is high in authority pages [4]. The algorithm starts with a root set R selected from the result list of a digital library search system. Starting with R, a set S is obtained keeping in mind that S is relatively small, rich in relevant pages about the query and contains most of the high authorities. HITS algorithm expands the root set R into a base set S by using the algorithm (see Fig. 5).
2. Iterative Step: This step finds hubs and authorities using the output of the sampling step. In this [14] each page is associated with two numbers: an authority weight a_i, and a hub weight h_i. Pages

with a higher $a_i$value are considered as better authorities and pages with a higher h_i value as better hubs.

Let A be the adjacency matrix of the graph S (output of sampling step), v denotes the authority weight vector and u denotes the hub weight vector.The weights a_i and h_i of all the nodes in S are dynamically updated by as follows:

$$v = (A^t \times u) \quad (6)$$
$$u = (A \times v) \quad (7)$$

If we consider that the initial weights of the nodes as

$u_0 = \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$, then $A^t \times \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$

After applying k steps we get the system:

$$v_k = (A^t \times A) \times v_{k-1} \quad (6a)$$
$$u_k = (A \times A^t) \times u_{k-1} \quad (7a)$$

Example Illustrating Working of HITS. The adjacency matrix of the graph is:

$$A = \begin{bmatrix}0&0&1&0&0&1\\0&0&1&0&1&1\\0&0&0&1&0&0\\0&0&0&0&0&0\\0&0&0&1&0&0\\0&0&1&0&0&0\end{bmatrix} \mathrm{A^t} = \begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\1&1&0&0&0&1\\0&0&1&0&1&0\\0&1&0&0&0&0\\1&1&0&0&0&0\end{bmatrix}$$

Assume the initial hub weight vector is: $u = \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}$

We compute the authority weight vector by:

$$v = (A^t \times u)$$

$$v = \begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\1&1&0&0&0&1\\0&0&1&0&1&0\\0&1&0&0&0&0\\1&1&0&0&0&0\end{bmatrix} \times \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix} = \begin{bmatrix}0\\0\\3\\2\\1\\2\end{bmatrix}$$

Then, the updated hub weight is

$$u = (A \times v)$$

$$= \begin{bmatrix}0&0&1&0&0&1\\0&0&1&0&1&1\\0&0&0&1&0&0\\0&0&0&0&0&0\\0&0&0&1&0&0\\0&0&1&0&0&0\end{bmatrix} \times \begin{bmatrix}0\\0\\3\\2\\1\\2\end{bmatrix} = \begin{bmatrix}5\\6\\2\\0\\2\\3\end{bmatrix}$$

By using (6a) and (7a), the authority weights and hub weights are iteratively calculated untill the values get converged as shown in Table 5.

By calculating the above equations iteratively, the page ranks of papers become:

HITS (C) > HITS (F) >HITS (E) > HITS (D) > (HITS (A), HITS (B))

Limitations of HITS. Following are some constraints of HITS algorithm [15]:

- *Distinction between Hubs and authorities:* It is not easy to distinguish between hubs and authorities because many sites act as hubs as well as authorities.
- *Topic drift*: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights.
- *Automatically generated links*: HITS gives equal importance to automatically generated links which may not have relevance for the user query.

2.6 PaperRank Algorithm

Zhang Guangqian [16] gave a new ranking method for publications ranking named PaperRank based on Google's PageRank. In this method, publication's rank score is determined on the basis of the reading value and other factor because it considers that the reading value of same papers may be different due to different readers. The reading value of a paper is related to its content, the periodical in which it was published, and the author of the paper. Thus, this method considers the factors such as content, journal, author, published time etc. in order to measure the reading value of papers. PaperRank of the publication p can be calculated as:

$$PaperRank = BR \times AR \times IF \times D \quad (8)$$

where BR represents the baserank, AR denotes the AuthorRank, IF denotes the impact factor of the journal in which it was published and D represents the published time of publication p. Various parameters used in the PaperRank calculation are explained below.

BaseRank

Thes BaseRank (BR) calculates the rank of the publication by using the PageRank algorithm. It considers the quoted time of cited publication and the importance of the citing publication. The BaseRank formula is given as:

$$BR(u) = c \sum_{v \in B(u)} \frac{BR(v)}{N_v} \quad (9)$$

where u represents a publication, B(u) is the set of citations that point to u, BR(u) and BR(v) are rank scores of publications u and v respectively, N_V denotes the number of publications cited by publication v (i.e. number of references), c is a factor used for normalization.

AuthorRank

This parameter assumes that if paper A cited by paper B and C at the same time, then, being cited by paper B authored by a popular and prestigious author contributes more to the Rank value of A than being cited by paper C with an unimportant author. Thus, it calculates the AuthorRank by considering the authors contribution in a certain academic field. The AuthorRank can be computer

by using an author citation network [17] which is a directed and weighted graph where nodes represent authors, edges represent citing relationships from author A to author B, and edge weights represent the number of times that author A cites author B. The AuthorRank can be calculated as:

$$AR(a) = \mathrm{d} \sum_{b \in B(\mathrm{a})} \frac{AR(\mathrm{b})}{N_{\mathrm{b}}} \qquad (10)$$

where a represents an author , AR(a) is the set of author's "citing" author a, AR(a) and AR(b) are AuthorRank of author's a and b respectively, N_b denotes the number of authors cited by author b, d is a normalization factor.

Impact Factor of Journal

This parameter assumes that if paper A is cited by paper B and C, and paper B was published in the core journal, and paper C was from unimportant journal, then the vote from paper B to A contributes more rank value to paper A than a vote from paper C to paper A. Thus, it considers the impact factor of journal to represent the weight of each journal. The formula for calculating the impact factor of the journal is defined as follow:

$$IF(j) = \frac{C}{A} \qquad (11)$$

where IF (j) represents the impact factor of journal j, A denotes the total number of papers published in journal j in the previous two years, and C denotes the quoted times of papers in the current year.

Published Time

This parameter considers the time of the publication. It assumes that sometimes a recently published publication having only one or two citations due to small time window may be important to reader in a certain field. Thus, it introduces the time factor D as follow:

$$D(p) = \frac{(t - min\{T(k)\} + 1)}{(max\{T(k)\} + 1)} \qquad (12)$$

where D(p) represents time factor of paper p, t is the year in which p was published, B(p) denotes the set of all the papers, T is a n*1 matrix composed by all the years in which all the papers were published, and n is the total number of all the papers.

Limitations of PaperRank. Researchers have shown that scientific publications naturally form a network on the basis of citation relationships. This algorithm can do well for the direct relationships i.e. citation and cited relationships, but it may not adequately reflect the lineage of scientific works. In such scenario, counting the indirect citation, indirect co-citation, and indirect co-reference, which are feasible in the Web environment may be considered.

4 A Comparison Study

By extensive study and literature analysis of some of the important web page ranking algorithms, it is concluded that each algorithm has some relative strengths and limitations. A detailed comparison of ranking algorithms studied is shown in Table 6. Comparison is done on the basis of some measures such as main techniques used, methodology, input parameters, relevancy, quality of results, importance and limitations.

Table 4 Iteration Method for HITS

Iterations	PR (A)		PR (B)		PR (C)		PR (D)		PR (E)		PR (F)	
	v	u	v	u	v	u	v	u	v	u	v	u
0	1	1	1	1	1	1	1	1	1	1	1	1
1	0	0.56	0	0.67	0.70	0.22	0.47	0	0.23	0.22	0.47	0.33
2	0	1.6	0	2.55	0.82	0.04	0.23	0	0.35	0.042	0.64	0.51
3	0	0.58	0	0.75	0.73	0.005	0.08	0	0.32	0.005	0.58	0.30
4	0	0.58	0	0.73	0.73	0.001	0.03	0	0.32	0.001	0.59	0.32
5	0	0.58	0	0.73	0.73	0.001	0.03	0	0.32	0.001	0.59	0.32

Table 6: COMPARISON OF RANKING ALGORITHMS

Algorithm → / Measures ↓	Citation Count	Time dependent Citation Count	PageRank	Popularity Weighted PageRank	HITS	PaperRank
Main Technique Used	Web Structure Mining	Web Structure Mining	Web Structure Mining	Web Structure Mining	Web Structure Mining, Web content Mining	Web Structure Mining, Web content Mining
Description	Results are sorted based on number of incoming citations.	Results are sorted based on time dynamics of the citation graph i.e. age of the citations	Computes scores at indexing time. Results are sorted by taking into account the importance of citing papers.	Results are sorted according to weighted citations as well as popularity factor of publication venue of paper.	Computes hub and authority scores of' n' highly relevant pages on the fly. Relevant as well as important pages are returned.	Computes new score of the top 'n' pages. Pages returned are more relevant.
I/P parameters	Backlinks	Backlinks, publishing time of paper	Backlinks	Backlinks, Publication venue	Backlinks, forward links, Content	Backlinks, authors, Impact factor, time of publish.
Working Levels	1	1	N*	N	< N	N
Complexity	O(N)	$O(N^2)$	O(log N)	O(MN)	<O(log N)	O(log N)
Relevancy	Less	Less(More than CC)	Less(more than CC, TDCC)	More (less than PaperRank)	More (less than PaperRank)	More
Quality of Results	Less	Higher than CC	Medium	Higher than PR	Less	High
Importance	Simplicity of computation. It is proven method which has been used for many years in scientometrics.	This method considers the freshness of citations by differentiating between the old and new citations.	It statistically analyses whole citation graph at once. It captures not just quantity, but also quality of citing papers.	This method overcome the limitation of impact factor and considers the popularity of publication venue.	This method provides good results by considering Hubs and Authorizes scores and also considers the content of the paper.	The pages are sorted according to the importance of citations, author journal.
Limitations	Unweighted ranking i.e. it treats all the citations equally and does not take into account time.	It does not take into consideration the different importance of each citation.	Results come at the time of indexing and not at the query time. Results are sorted based on importance of citations.	It does not take into account the time of publication and also does not differentiate between the popular and prestigious authors.	Topic drift and efficiency problem.	Extra calculations to find the author ranking and time impact of citations.

*N: NUMBER OF PAPERS, M: AVERAGE CITATIONS TO A PAPER

REFERENCES

[1] Naresh Barsagade, "Web Usage Mining And Pattern Discovery: A Survey Paper", CSE 8331,2003

[2] M. Krishnamurthy, "Open access, open source and digital libraries: A current trend in university libraries around the world", A General Review, Emerald Group Publishing Limited

[3] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, pp. 558-567, 1997

[4] N. Duhan, A.K. Sharma and K.K. Bhatia, "Page Ranking Algorithms: A Survey", Proceedings of the IEEE International Conference on Advance Computing, 2009

[5] R. Kosala, and H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, vol. 2, no. 1, pp. 1-15, 2000

[6] Joeran Beel, Bela Gipp, "Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study)", In Rcis 2009: Proceedings of The IEEE International Conference on Research Challenges In Information Science, 2009

[7] L. Marian, M. Rajman, "Ranking Scientific Publications Based on Their Citation Graph", Master Thesis, CERN-THESIS, 2009

[8] L. Marian, J. Yves LeMeur, M. Rajman, M. Vesely, Citation Graph Based Ranking in Invenio", ECDL, pp. 236-247, 2010

[9] L. Page, S. Brin, R. Motwani, and T. Winograd.P, "The Pagerank Citation Ranking: Bringing order to the Web. Technical report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999

[10] Sergey Brin and Larry Page, "The anatomy of a Large-scale Hypertextual Web Search Engine", In Proceedings of the Seventh International World Wide Web Conference, 1998

[11] A. Sidiropoulos, Y. Manolopoulos, "Generalized Comparison of Graph-based Ranking Algorithms for Publications and Authors", Journal of Systems and Software archive, vol. 79, issue 12, pp. 1679-1700, 2006Y. Sun, C. L. Giles, "Popularity Weighted Ranking for Academic Digital Libraries", In 29th ECIR, pp. 605-612, 2007

[12] Kleinberg J., "Authorative Sources in a Hyperlinked Environment", Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998

[13] Math Explorer's Club, "The Mathematics of Web Search", http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

[14] Longzhuang Li, Yi Shang, Wei Zhang, " Improvement of HITS based Algorithms on Web Documents", WWW2002,Honolulu Hawaii USA, 2002

[15] Zhang Guangqian, Liu Xin, "Study on the Method of Ranking Scientific Papers", In The International Conference on E-Business and E-Government, Guangzhou China, pp. 3062-3066, 2010

[16] Ding Y., "Applying Weighted PageRank to Author Citation Networks", Journal of the American Society for Information Science and Technology, pp. 236-245,