Volume 4, Issue 6, June 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Survey on Page Ranking Algorithms for Digital Libraries

Deepti Kapila
Research Scholar in computer science department
RIMT-IET, Punjab,
India

Prof. Charanjit Singh
Prof. In computer science department
RIMT-IET, Punjab,
India

Abstract— Digital libraries have become an important tool for searching the research papers for scientific purpose. The growth of digital libraries is increasing day by day. To make the search better, the ranking of digital libraries should be done properly. The rank of research paper in digital libraries depends upon many factors like citations to paper, content, authors, relevancy and publications of the paper etc. Based upon these parameters, different ranking algorithms have been developed. In this paper, the ranking algorithms is discussed, which considers important factors like citations to the paper and the relevancy of the content with the query, input and output parameters, publications of the paper etc.

Keywords— digital library, page rank, citation, web mining, search engine

I. INTRODUCTION

Number of research papers is published every year and these papers span various fields of research. For a new researcher, it becomes a very difficult task to go through the entire repository of research papers in order to determine the important ones. There can be several ways of determining whether a research paper is important depending on the field of work, conference of publication, etc. To make the search easier for user, digital library contains huge set of research papers according to user requirements. But, it is important to rank the research papers in digital libraries so that users could find research paper according to their interest. Sometimes, user fires a query on search engine, but does not get relevant result. For more accuracy and relevancy the ranking is important. The digital library is important tool where user can get scientific literature. Digital libraries have been introduced to make retrieval mechanism more effective and relevant for researchers or users. The digital libraries are the part of electronic source where collection of all research papers and journals are placed according to their relevancy, conference and publication etc.

II. ARCHITECTURE OF DIGITAL LIBRARY

A digital library is an integrated set of services for capturing, storing, searching, protecting and retrieving information, which provides coherent organization and convenient access to typically large amounts of digital information. It is an electronic resource where user gets research papers, articles, journal and material related to research work or survey. Digital libraries are important tool now days, and the growth of digital libraries is increasing rapidly. The main component of the digital library search system is a crawler and this component crawler traverses the hypertext structure in the web, it 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. Thus, when user searches for a new topic, a new instance of the agent is created for that particular topic which locates and downloads postscript files. These downloaded files are passed through the document parsing sub agent who extracts the semantic features 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. The architecture of a digital library search engine is shown in Fig.1

III. WEB MINING AND ITS CATEGORIES

World Wide Web (WWW) is an architectural framework for accessing linked documents spread over millions of machines all over the internet. The popularity of WWW is largely dependent on the search engines. Search engines are the gateways to the huge information repository at the internet. Now a user can access the information at search engine very easily.

I. *Web Mining:*

Extraction of interesting information or patterns from large databases is called Data Mining. The web mining is the main Part of data mining techniques to discover and retrieve useful information from WWW. It is the mechanism to Classify contents and used to retrieve information from the WWW documents.

II. *Categories of Web Mining:*

Web mining consists of three main categories: web content mining, web structure mining, and web usage mining.

a) **Web Content Mining:** It is the process of scanning and mining the text, pictures and graphs of web pages. It is used to determine the relevance of the content of the web pages according to the search query.

b) **Web Structure Mining:** Extraction of interesting information or patterns from large databases is called Data Mining. Web Mining is the application of data mining techniques to discover and retrieve useful information from the WWW documents and services.

c) **Web Usage Mining:** It extracts data stored in server access logs, referrer logs, agent logs, client-side cookies, user profile and Meta data.

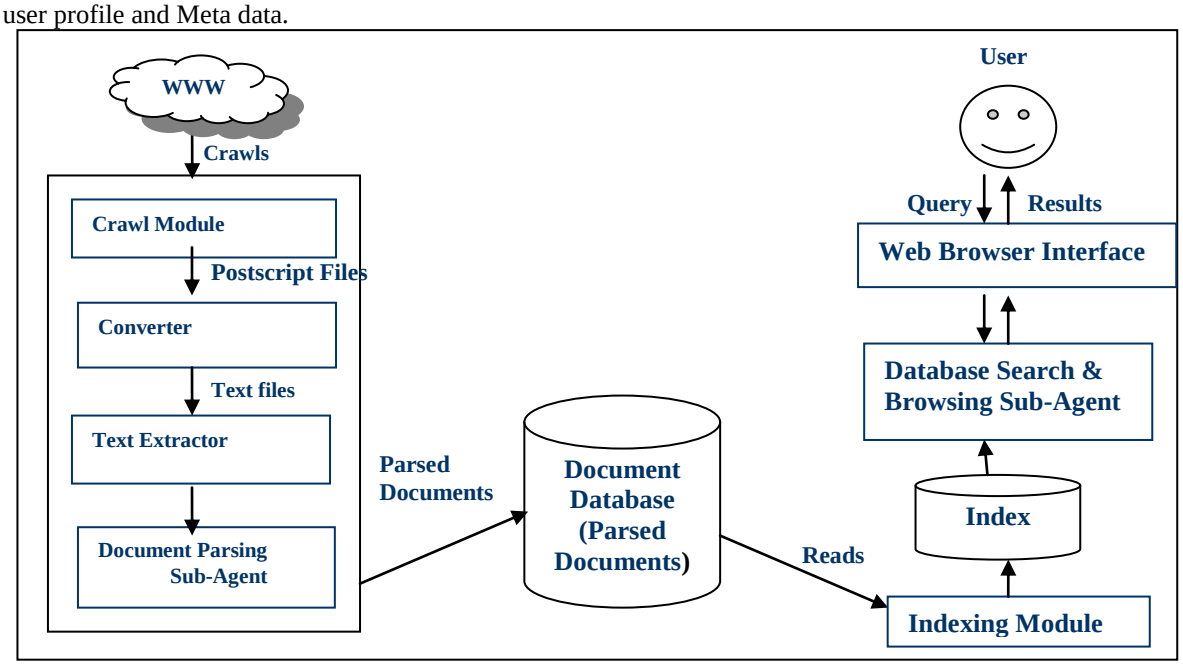

Figure 1: Architecture of Digital Library

Web Usage Mining (WUM) is a process of identifying the browsing patterns by analysing the user's navigational behaviour while surfing on the web. The categorization of web mining is divided in to three categories which are shown in figure 2. Web content mining is further divided in to two parts: web page content mining, web search results mining.

d) **Web Page Content Mining:** It displays the web pages according to relevant contents at the search engine.

e) **Web Search results Mining:** It displays the results according to user puts a query at search engine and it also extracts useful information at web pages.

Web usage mining is further divided in to two parts: General access pattern tracking, Customized usage tracking.

f) **General Access Pattern Tracking:** It displays the useful information in patterns such as links, graphs, venue, documents etc.

g) **Customized Usage Access:** The part of web usage mining where user can get extract information in the form of server logs or browser logs.

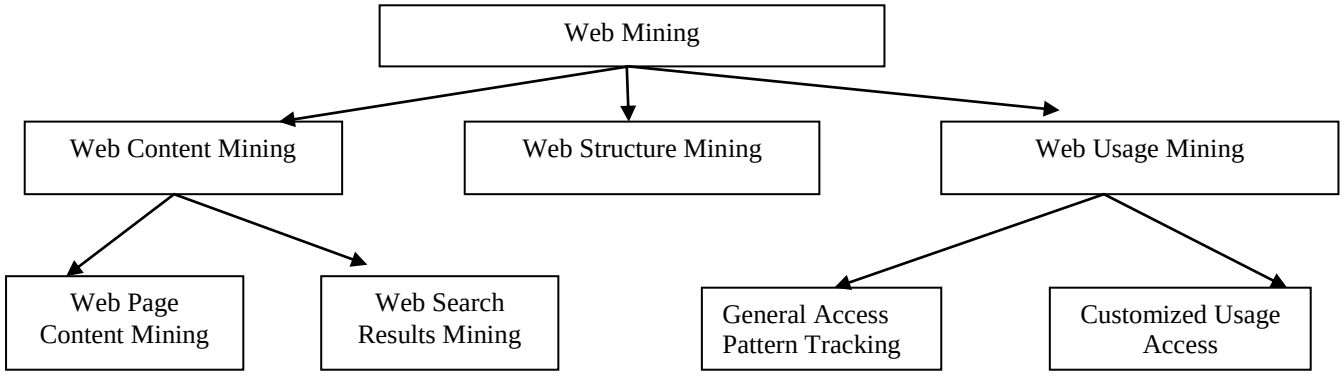

Figure 2: Categories of Web Mining

IV. PAGE RANKING ALGORITHMS

Ranking algorithms are used by the various search engines. There are various algorithms by which user can get relevant results at search engines. Some ranking algorithms are based on content in the documents, some depend on the link structure of the documents and some are combination of both. If search results are not displayed according to the user interest then search engine will lose its popularity. So the ranking algorithms become very important. Some of the page ranking algorithms are discussed as below:

V. PAGE RANK ALGORITHM

Page rank algorithm is proposed by Brin and Page during their PhD at Stanford University based on the citation analysis. Page rank algorithm is used by famous search engine Google, where the most important web pages are displayed at the top and less relevant pages at the bottom. The Brin and Page applied the citation analysis in web search by treating the incoming links as citations to the web page. Page Rank provides a more advanced way to compute the importance or relevance of web page than simplify the number of pages that are linking to it. For example, if a web page has a link of the Yahoo! home page, it may be just one link but it is a very important one. This page should be ranked higher than many pages with more links but from obscure places. Page rank is an attempt to see how good an approximation of importance can be obtained just from the link structure. The Page rank algorithm provides a more sophisticated method for doing citation counting. The reason that Page rank is interesting is that there are many cases where simple citation counting does not correspond to our common sense notion of importance. Based on the discussion above, is it possible to give the following intuitive description of Page rank: a page has high rank if the sum of the ranks of its back links is high. This covers both the case when a page has many back links and when a page has just a few (but highly ranked) back links.

Page rank is calculated by:

PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn)) (i)

Where PR (A) is the Page Rank of page A, PR (Ti) is the Page Rank of pages Ti which link to page A, C (Ti) is the number of outbound links on page Ti and d is a damping factor which can be set between 0 and 1.Page Rank does not rank web sites as a whole, but is determined for each page individually. Further, the Page Rank of page A is recursively defined by the Page Ranks of those pages which link to page A. We describe a small web consisting of three pages A, B and C, whereby page A links to the pages B and C, page B links to page C and page C links to page A. According to Page and Brin, the damping factor d is usually set to 0.85, but to keep the calculation simple we set it to 0.5. The exact value of the damping factor d admittedly has effects on Page Rank, but it does not influence the fundamental principles of Page Rank.

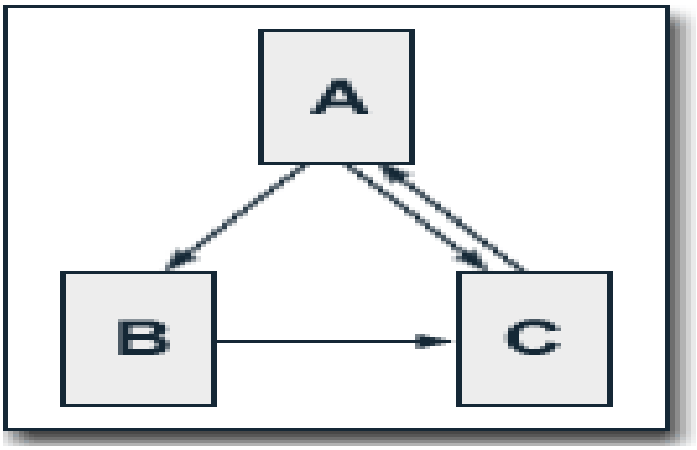

Figure 3: Incoming links in Page Rank

The Figure 3 describes the page rank incoming links from where page rank values for single page can be calculated. we get the following equations for the Page Rank calculation:

PR(A) = 0.5 + 0.5 PR(C)

PR(B) = 0.5 + 0.5 (PR(A) / 2)

PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

These equations can easily be solved. We get the following Page Rank values for the single pages:

PR(A) = 14/13 = 1.07692308

PR(B) = 10/13 = 0.76923077

PR(C) = 15/13 = 1.15384615

Advantages of Page Rank:

I) The important web pages are kept at the top and irrelevant pages are kept at the bottom.

II) It is representation of web structure mining which extracts useful information in terms of relevant web pages.

III) It calculates the important web pages by incoming and back links and the representation is simple i.e. in graph and linked databases.

VI. CITATION COUNT ALGORITHM

Citation count is one of the most popular used ranking algorithms. It is used to measure a scientist's reputation, as named Citation Count It takes back links into account to order the publications. Thus, a publication obtains a high rank if the number of its back links is high. Citation Count is calculated by:

CC=|Ii| (ii)

Where CCi represents the citation count of publication i and |Ii| denotes the number of citations (in-degree) of the publication i . We take a citation graph as example to calculate its back links as the graph is shown in figure 4.

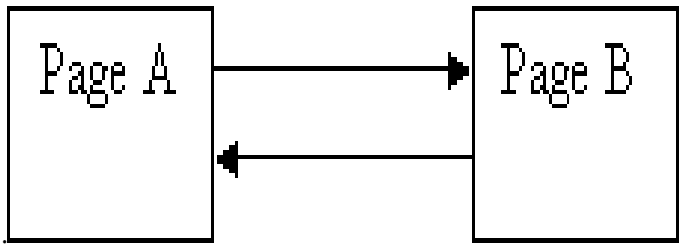

Figure 4: Network of Citation Graph Representation

Let us take the example of simplest network to measure the citations where two pages are considered, each pointing to each other page. Each page has its one outgoing link as shown in figure 3. So, calculation of citations is:

C (A) =1 and C (B) =1

Let us take another example to measure the citations as shown in Fig. 5 and Table 1 where A, B , C, D, E and F are six publications.

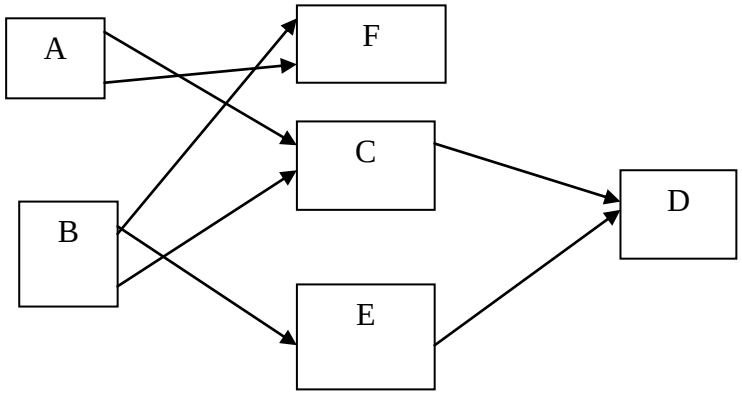

Figure 5: Citation Count Network Measurement Graph

The Citation Count for publications A, B, C, D, E and F can be calculated by using equation (ii):

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))

The above example of citation graph states that if a publication has more publications, then it becomes important.

Publication	Publication year
A	2013
B	2008
C	1998
D	1980
E	2007
F	2000

Table 1: For data of Citation graph

The above example shows how citations are measured by citation network graph.

Advantages and limitations of Citation Count Algorithm:

I) It measures the back links but**,** 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.

II) It measures the reputation of scientist but**,** it does not take into consideration different characteristics of the citations, like their publication date.

VII. HITS(HYPERLINKED INDUCED TOPIC SEARCH) ALGORITHM

HITS algorithm is proposed by Kleinberg in 1988. HITS algorithm identifies two different types of web pages called hubs and authorities. Authorities are pages having important contents. Hubs are pages that act as resources list, guiding users to authorities. Hyperlink induced topic search(HITS) assumes that for every query given by the user, there is a set of authority pages that are relevant and popular focusing on the query and a set of hub pages that contain useful links to relevant pages including links to many authorities .

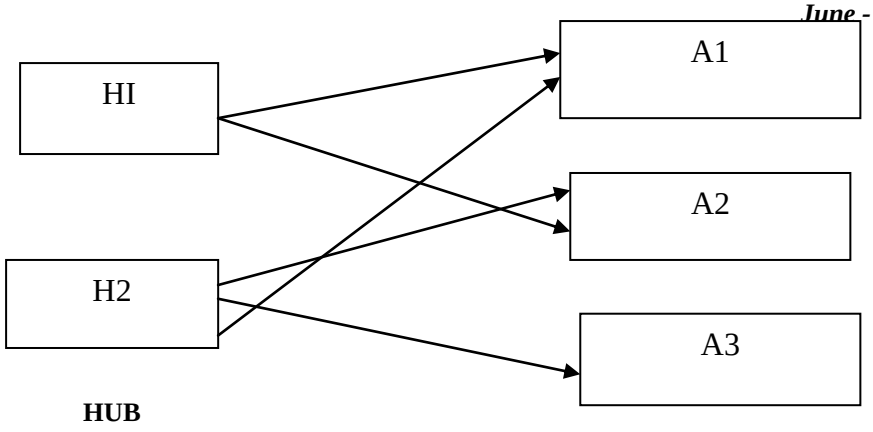

Figure 6: Representation of Hubs and Authorities

The figure 6 shows the number of page p provides a link to page q, then p confers some authority on page q.

Advantages and limitations of HITS Algorithm:

I) It contains useful links to relevant pages including links to many authorities but, high rank value is given to some popular website that is not highly relevant to the given query.

II) It is simplest form to calculate the rank by hubs and authorities but, Topic drift occurs when the hub has multiple topics as equivalent weights are given to all the out links of a hub page.

VIII. DISTANCE RANK ALGORITHM

A distance rank algorithm is proposed by Ali Mohammad Zareh Bidoki and Naseer Yazdani. The main goal of distance rank algorithm is computed on the basis of the shortest logarithmic distance between two pages and ranked according to them so that the page with smaller distance to assigned a higher rank. The advantage of this algorithm is that, it can fin pages faster with high quality and more quickly. This algorithm uses the page rank algorithm properties. This ranking algorithm is based on recursive method .It uses the property of page rank that the page has a higher rank value if it has more incoming links on a page.

Advantages and limitations of Distance Rank Algorithm:

I) It calculates the shortest distances between two pages but, it does not satisfy the content relevancy of pages.

II) This algorithm uses the property of page rank algorithm and calculation speed to find relevant pages is also faster but, limited in use and takes in to account more calculations. So, it is complicated to implement by calculating all the factors.

IX. EIGENRUMOR RANK ALGORITHM

This algorithm is proposed by Ko Fujimura. This algorithm ranks each blog entry on the basis of weighting the hub and authority scores of bloggers. So, this algorithm enables a higher score to be assigned to a blog entry entered by a good blogger. The rank scores of blog entries are decided by the page rank algorithm is often low so it cannot allow blog entries to be provided by rank score according to their importance. So, to resolve the issue, an Eigenrumor algorithm is proposed for ranking the blogs.

Advantages and limitations of Eigenrumor Algorithm:

I) It calculates ranking on the basis of each blog entry, which is part of HITS algorithm but, rank score is decided by page rank.

II) It is used to rank the blogs but, calculation of hubs and authority is difficult to calculate.

X. COMPARISON STUDY

Web services has vast amount of information, where user can access the information by firing the query at the search engine. But, it is difficult to access the relevant information by the users and sometimes, user does not get relevant result as the result given by search engine is relevant to one user but less relevant to another user. So, the ranking of documents in web services are required.

The available algorithms highlight several differences in the basic concepts used in each algorithm. The web mining technique is categorized into different types where, HITS is an iterative algorithm like page rank because of based on link structure of documents on the web based on analysis, the comparison of some of various web page ranking algorithms is shown in table 2.

This comparison is done on the basis of some parameters such as relevancy, quality of results, basic technique, and importance.

Table 2: Comparison study of Page Rank Algorithms

Algorithm name:	Page rank	Citation count	HITS	Distance rank	Eigenrumor
Main techniques:	Web Structure Mining(WSM)	Web Structure Mining(WSM)	WSM and WCM	WSM	WCM
Working process:	It calculates the score at result time. Results are sorted by taking into account the importance of citing papers.	It calculates the results based on number of incoming citations.	'n' highly relevant pages are calculated and find values on the fly.	It calculates the minimum average distance between two pages and more pages.	It uses the adjacency matrix which is constructed from agent to object link not page to page.
I/P parameters:	Back links	Back links	Back links and forward links and content	Back links	Agent/Object
Complexity:	N^*	1	<O(log N)	O(log N)	<O(log N)
Limitations:	Query independent	Query dependent	Topic drift and efficiency problem	Needs to work along with PR	Used for blog ranking
Search engine:	Used in Google	Used in Research model	Used in IBM	Used in Research model	Used in Research model

REFERENCES

[1] D.Neelam, Sharma. A.K, Bhatia.Komal,"Page ranking algorithms", IEEE, Vol.4, Issue4, April 2013, pp.2811-2818.

[2] G.Sumita, D.Neelam and B.Poonam,"A comparative study of page ranking algorithms for online digital libraries", International journal of scientific &.engg. Research, vol.4, issue4, April 2013, pp.1225-1233.

[3] M.Debajyoti, B.Pradipta and K.Young-chon,"A syntactic classification based web page ranking algorithm", 6th International workshop on MSPT proceedings, 2006, pp.83-92.

[4] D.Sujatha, M.Prasenjit and D.Lee,"Ranking authors in digital libraries", IEEE joint conference in digital libraries, 2011, pp.251-254.

[5] Taneja, H., Gupta, R., "Web Information Retrieval using Query Independent Page Rank Algorithm", International Conference on Advances in Computer Engineering, Published in IEEE, Print ISBN No: 978-0-7695-4058-0, 2010, pp. 178-182.

[6] Naresh Barraged; Web Usage Mining and Pattern Discovery: A Survey Paper. CSE 8331, 2003.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd; The page rank citation ranking bringing order to the web. technical report, computer science department, Stanford university; 1998.

[8] RankDex; the RankDex search engine. available online at http://rankdex.gari.com/

