Impact of Page rank and Citation Count Algorithm for Digital Libraries - A Review

Deepti Kapila
Assistant professor
Computer science department
LCET, Katani Kalan

Charanjit Singh
Assistant professor
Computer science department
RIMT, Mandi Gobindgarh

Shikha Singla
Assistant professor
Computer science department
LCET, Katani Kalan

ABSTRACT
With the huge number of web pages that exist today, search engine plays an important role in the current Internet. But even if they allow finding relevant pages for any search topic, now a day the number of results returned is often too big those are very difficult to manage by the users. Moreover, the needs of the users vary, so that what may be important for one user may be completely irrelevant for another user. The role of ranking algorithms is thus important, in this select the pages that are most likely be able to satisfy the user's needs, and bring them in the top positions. In this survey the most popular algorithms used today by the search engines i.e. Page Rank and Citation count are introduced which are based on web structure mining technique .The comparison study of page rank and citation count algorithm is introduced to bring the difference between these two algorithms.

Keywords
Web Mining, Digital library, Page rank, Citation

1. INTRODUCTION
Thousands of research papers are 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 blogs of research papers in order to determine the important ones. The term important is subjective but it can be assured that a research paper that is popular will be important in most cases. There can be several ways of determining whether a research paper is important depending on the field of work, conference of publication, etc.

An efficient ranking algorithm is important in any information retrieval system. In a web search engine, due to the dimensions of the current web, and the special needs of the users, its role becomes critical. 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 to a 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. As user wants to get the results in short span, it is necessary to rank the pages which are relevant according to the user's input query. To overcome this problem, 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.

2. DIGITAL LIBRARY
A digital library 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. It is an electronic resource where user gets research papers, articles, journal and material related to research work or survey. Now a day, digital libraries are experiencing rapid growth with respect to both the amount and richness of available digital content. The main component of the digital library search 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. 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.

3. WEB 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. We divide the web mining into three categories: WSM, WCM, and WUM.

3.1 Web structure mining (WSM)
Web Structure Mining (WSM) means to discover the useful knowledge from the structure of hyperlinks 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.

3.2 Web content mining (WCM)

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.

3.3 Web usage mining (WUM)

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.

The Categories of web mining technique is shown in Fig. 2. Where web mining is divided into three categories and both the algorithms, page rank and citation count are based on web structure mining.

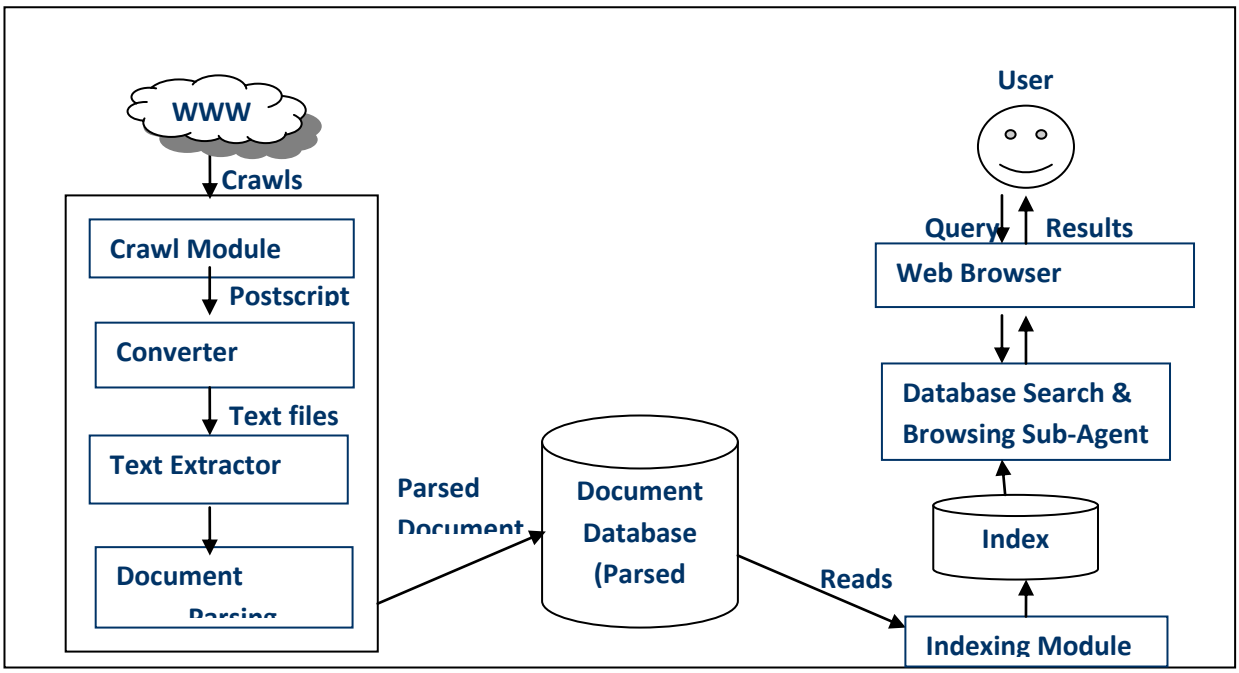

Fig 1: The Architecture of Digital Library Search Engine

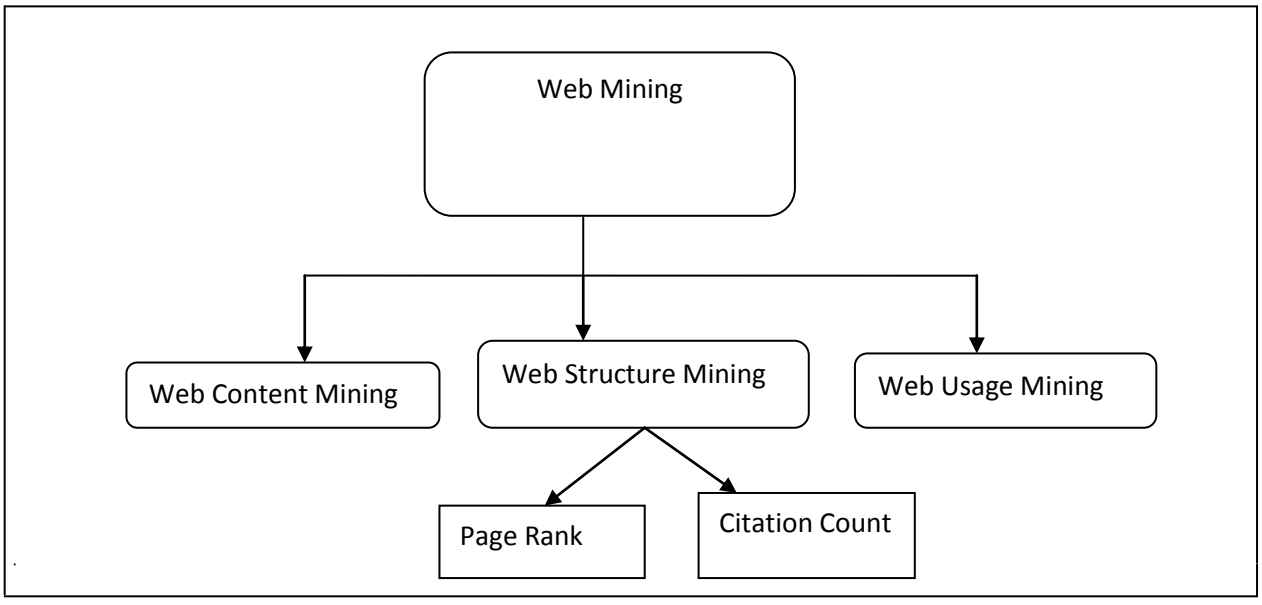

Fig 2: The Categories of Web Mining Technique

4. PAGE RANK ALGORITHM

Brin and page developed Page Rank Algorithm during their PhD at Stanford University based in the citation analysis. Page Rank algorithm is used by the famous search engine, Google. They 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.

It is given by:

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

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.

4.1 Calculation of Page rank

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. So, 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

The fig: 3 describe the page rank incoming links from where page rank values for single page can be calculated.

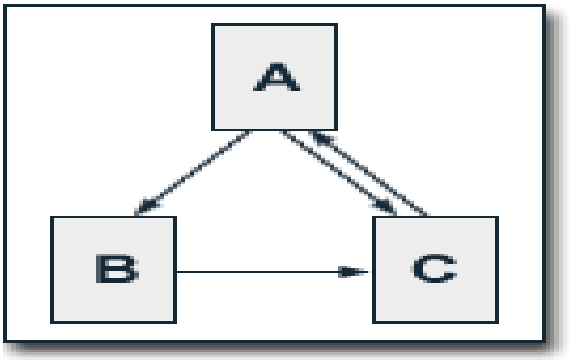

Fig 3: Page Rank Back Links (Incoming Links)

5. CITATION COUNT ALGORITHM

This is one of the most frequent used ranking algorithms for measuring 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 defined in (1):

$$CC=|Ii| \quad (1)$$

where CCi represents the citation count of publication i and |Ii| denotes the number of citations (in-degree) of the publication i.

5.1 Calculation of citations

let us take an example of citation graph as shown in Fig. 4 and Table 1 where A, B , C, D, E and F are six publications.

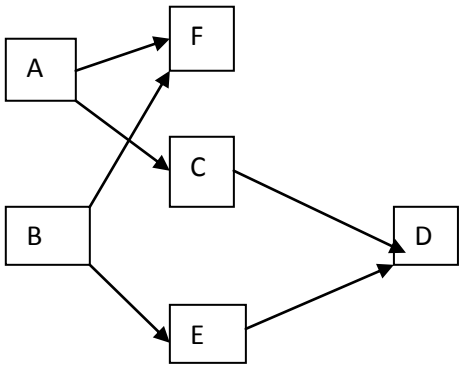

Fig: 4 Example of Citation Count Graph

The citation count can be calculated by its incoming links as the example shows different incoming links from various publications of A, B, C, D, E and F.

The example of citation graph shows publication links and table 1 show its publication links with year of publication**.**

Table 1 for data of Citation graph

Publications	Publication year
A	2013
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 equation (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))
The above example of citation graph states that if a publication has
More number of citations to it then publication becomes important.
So, publication gets high rank if the number of its back links is high.

6. A COMPARISON STUDY

By extensive study and literature analysis of page rank and citation count web page algorithms, it is concluded that both algorithm has some relative strengths and limitations.

A comparison study of both algorithms is shown in Table 2.The page rank and citation count algorithms are used for ranking the research papers in digital libraries. The importance of both algorithms is based on web structure mining where results are sorted according to citing papers and incoming links. This comparison is done on the basis of some measures such as main techniques used, methodology, input parameters, relevancy, and quality of results, importance and limitations.

7. CONCLUSION

Web Mining is used to extract the useful information from very large amount of web data. The usual search engines usually results in a large number of web pages in response to user's queries, while the user always want to get result in short time. So, Page ranking and citation count algorithm, which are an applications of web mining, play a vital role in making the user search navigation easier in the results of a search engine .The page rank and citation count algorithms are very useful for digital libraries as user wants to get the results in short time. Both algorithms are based on some measurements according to their importance, relevant and contents. Depending on the technique, both algorithms have similarities and dissimilarities also. As a future guidance, the algorithm which equally considers the prestigious authors as well as the importance of a page should be developed so that the quality of results can be improved.

TABLE 2: A Comparison Study of Page Rank and Citation Count Algorithm

Algorithm Name:	**Page Rank**	**Citation Count**
Main Technique used:	Web Structure Mining(WSM)	Web Structure Mining(WSM)
Description:	It computes the score at result time. Results are sorted by taking into account the importance of citing papers.	It computes the results based on number of incoming citations.
I/O Parameters:	Backlinks	Backlinks
Working Levels:	N*	1
Complexity:	O(log N)	O(N)
Relevancy:	Less(but more than CC)	Less
Quality of Result:	Medium	Less
Advantages:	It analyses whole citation graph at once. It captures not only quality, but also quality of citing papers.	It is simple method which has been used for many years for computation of back links.
Limitations:	Results come at the time of indexing and not the query time.	It treats all the citations equally and does not take into account of time.

8. REFERENCES

[1] Sumita Gupta, Neelam Duhan, Poonam Bansal; A comparative study of page ranking algorithms for online digital library. International Journal of Scientific & Engineering Research, Volume 4, Issue 4, April 2013.

[2] 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.

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

[4] S.Pratap, S.Kumar and P.Vikram,"An efficient algorithm for ranking research papers based on citation

[5] B.J.Jansen, A.Spink, J.Bateman, T.Saracevic; Real life Information Retrieval: A study of user queries on the web. ACM SIGIR Forum, 1998. Network,"IEEE conference in digital libraries, june 2011, pp.88-95.

[6] Craig Silverstein, Monika henginger, hannes marais and Michael moricz; Analysis of very large altavista query log. Tech. Report 1998-014, Digital SRC, 1998.

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

[8] M. Henginger; Hyperlink analysis on the web. 2003; available online at http://www.cad.eecs.berkeley.edu/tah/170/Notes/170-google.ppt

[9] A. Arasu, J.Cho, H. Garcia-Molina, A. Paepcke and S. Raghavan; searching the web. ACM transactions on internet technology; 2001.

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

[11] 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..

[12] A.Broder; Web Searching Technology Overview. Advanced school and Workshop on Models and Algorithms for the World Wide Web, 2002.