Page Ranking Algorithms in Online Digital Libraries: A Survey

[1]Sumita Gupta, [2]Neelam Duhan, [3]Poonam Bansal, [4] Jigyasa Sidhu
[1,2] YMCA University of Science & Technology, Faridabad, India, [3] Surajmal Institute of Technology, Delhi, India, [4] Amity School of Engineering & Technology, Noida,India
sumitagoyal@gmail.com, neelam_duhan@rediffmail.com, pbansal89@yahoo.co.in, jigyasa.sidhu@gmail.com

***Abstract*—With the exponential growth of academic digital libraries, ranking has become an extremely challenging task. When a researcher tries to retrieve relevant scientific literature, or perform a literature review based upon which to build their research, then ranking of search results of a user query plays an important role. Ranking provides an order so that users can easily navigate through the search results and find the desired information content. The various ranking algorithms have been proposed based upon many factors like citations to publications, content similarity, annotations etc. This paper presents an outline regarding various page ranking algorithms for academic digital libraries and highlights comparison of these algorithms in context of performance. This comparative analysis encourages for further required improvement in the related field.**

IndexTerms—WWW, Digital library, Page ranking, Search engine, Web Mining

I. Introduction

With exponentially increasing the size of information available on the World Wide Web (WWW), it is becoming difficult to trace the relevant information available on the Web and satisfy the user needs [1]. For this purpose, many advanced web searching and mining techniques have been recently developed to find, extract, filter or evaluate the relevant information as per the user need. In spite of advances in search engine technologies, there still occur situations 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. 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 [3]. 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.

In this paper, a survey of some prevalent paper 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, a detailed overview of some page ranking algorithms with their advantages and limitations has been discussed. Section III presents an extensive comparison study based on many parameters like main technique used, methodology, quality of results, complexity etc.. Finally in Section IV, conclusion is drawn with a light on future suggestions.

II. PAGE RANKING ALGORITHMS IN DIGITAL LIBRARIES

With the extensive growth of digital content available on WWW, the number of user and their queries to access the required content is also increasing. Therefore, the main challenge in front of search engines is to efficiently extract the relevant scientific work and order the search result list. To represent the documents in an ordered manner, various page ranking methods are employed utilizing different schemes to calculate relevance and importance of a document based on different parameters. Some of the common page ranking algorithms for online digital libraries have been discussed here as follows.

A. PageRank Algorithm

Surgey Brin and Larry Page [4][5] at Stanford University proposed a new algorithm named PageRank (PR) algorithm. It has become the most commonly used ranking algorithm. It has encouraged and laid the foundation for many other ranking schemes. This algorithm considers the link structure of the research papers for ranking the result list of a search engine. In other words, the rank of a paper is calculated by considering incoming links as well as outgoing links of a research papers. This algorithm states that if an incoming link comes from an important paper then, this link is given higher weightage than those coming from non-important paper. Using this algorithm, the PageRank of a paper p can be calculated as:

$$PR(p) = (1-d) + d\sum_{q\in B(p)} \frac{PR(q)}{N_q} \qquad (1)$$

where p represents a paper, B(p) is the set of papers that points to p, PR (p) and PR (q) are rank scores of papers p and q respectively, N_q denotes the number of outgoing links of paper q, and d is the damping factor.Damping factor determines the probability that the user will continue clicking the links. The value of damping factor ranges between 0-1 and the best results are found at d=0 .85.

978-1-4799-6896-1/14/$31.00 ©2014 IEEE

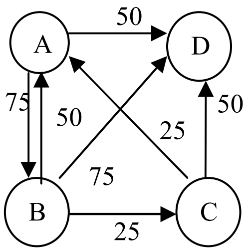

Fig 1. Citation Graph

Table 1. Iteration Method for PR

PR(A)	PR(B)	PR(C)	PR(D)
1	1	1	1
0.858	0.514	0.295	0.784
0.421	0.328	0.242	0.524
0.345	0.296	0.233	0.479
0.332	0.291	0.232	0.472
0.331	0.290	0.232	0.471
0.330	0.290	0.232	0.471
0.330	0.290	0.232	0.471

Illustration of PageRank Algorithm

Consider an example of the citation graph as shown in Fig. 1 where A, B, C and D represents the four publications. PageRank algorithm is applied to calculate the page rank of every paper in the graph. The value of damping factor is assumed to be 0.85.

The PageRank for papers can be calculated by using (1):

$$PR(A) = (1-0.85) + 0.85\left[\frac{PR(B)}{3} + \frac{PR(C)}{2}\right] \quad (1a)$$

$$PR(B) = (1-0.85) + 0.85\left[\frac{PR(A)}{2}\right] \quad (1b)$$

$$PR(C) = (1-0.85) + 0.85\left[\frac{PR(B)}{3}\right] \quad (1c)$$

$$PR(D) = (1-0.85) + 0.85\left[\frac{PR(A)}{2} + \frac{PR(B)}{3} + \frac{PR(C)}{2}\right] \quad (1d)$$

The equations (1a), (1b), (1c) & (1d) are solved iteratively until the page ranks get converged. The final values drawn are summarized in the Table 1.

Thus, the final results conclude that the rank score of page D is the highest among the three pages A, B, and C. The order obtained is:

PR (D) > PR (A) > PR (B) > PR (C).

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. On the other hand, there are some shortcomings of this ranking method also [3][6]:

- 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 rank scores of the publications that point to it and less affected by the number of citations.
- Pagerank algorithm only depends on the link structure of citation graph instead of the query.
- This method emphasizes more to the old publications as compared to newer publications. Thus, when user wants to search for latest publication, then the returned result list do not necessarily reflect the real important and relevant information to the user.
- PageRank gives high score to a publication, if it contained a 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.

B. Popularity and Similarity Based Page Rank Algorithm (PSPR)

Phyu Thwe [8] proposed a Page Rank like algorithm for conducting a web page access prediction named as *Popularity and Similarity Based Page Rank Algorithm (PSPR)*. This method highlights an improvement in the prediction of web page access by a user [9]. It is based on Web Usage Mining and processes the web server log files to analyze the user's browsing pattern for predicting user's next click. This method ranks the result list of a search engine by taking into consideration the popularity and similarity among web pages as well as the user's navigation behavior pattern.

PSPR functions in two major steps:

1. *Build Markov Model*: In this step, Markov models [8] [9] is used for predicting the behavior of a web user. It is the most widely used web usage mining algorithm for modeling sequences or processes of browsing behavior of a user using finite- state structure. This model takes web pages in the sequence accessed by a user as input parameter and output a model that predicts the user next access/click. Let assume P be a set of web pages in a web site, W be a user session W of a website. P can be written as $P= \{p_1, p_2 ... p_n\}$.Then, the probability of visiting the next page p by the user is denoted by conditional probability $P = (p_i|W)$. Assuming that i number of pages has already been visited by the user. From here it can be said that the prediction of next page access does not depend on all the pages in a web session rather can be restricted to small number of k pages. The number k also marks the order of the Markov model. Thus, it can be judged that the web page p_{i+1} will be accessed next using (2),

$$P_{i+1} = argmax_{p \in P}\{P(P_{i+1} = p|p_i, p_{i+1}, \dots p_{i-(k-1)})\} \quad (2)$$

2. *Similarity Calculation*: The popularity of page and transitions plus similarity among the web pages is determined to calculate the importance of the web pages. Similarity is computed based on the contents of the page URL. Following steps are taken in this method:
    - Select the URLs of the two pages so as to calculate similarity among them.
    - The URLs are sorted in a string array being separated by a special character '/' and their length is calculated.
    - Weights are assigned to each array starting from the longest array to the smallest one.

Table 2. A web session for a website

Session ID	Transitions
ID1	C, B, A
ID2	D, E, B, A, E, D
ID3	A, D, E, B, D
ID4	A, D, B, E, C

Table 4. 2nd Order Transition Probability Matrix

	A	B	C	D	E
{A,D}	0	1	0	0	1
{A,E}	0	0	0	1	0
{B,A}	0	0	0	0	1
{B,D}	0	0	0	0	0
{B,E}	0	0	1	0	0
{C,B}	1	0	0	0	0
{D,B}	0	0	0	0	1
{D,E}	0	2	0	0	0
{E,B}	1	0	0	1	0
{E,C}	0	0	0	0	0
{E,D}	0	0	0	0	0

Table 3. Ist Order Transition Probability Matrix

	A	**B**	**C**	**D**	**E**
s1=A	0	0	0	2	1
s2=B	2	0	0	1	1
s3=C	0	1	0	0	0
s4=D	0	1	0	0	2
s5=E	0	2	1	1	0

Table 5. An Example of Similarity Calculation

Page	Page URL
A	/project/creators/order-23/madeasy.html
B	/project/creators/artificial/madeasy.html

- The matching substrings are identified and their corresponding weights are added and the sum is divided by the total weight to give the similarity measure between the two.

The similarity of two web pages lies between 0.0 and 1.0. If similarity comes out to be 1, it indicates that the two web pages are exactly same. But, if it comes out to be 0, then it is concluded that the web pages are totally different. But this method fails to predict directly one more step ahead.

Example Illustrating Working of PSPR:

*Building Markov model:*Let us assume a sample web session of any website as shown in Table 2 for building Markov model where Session ID represents the different users and Transitions represents the sequence of pages access by particular user.

Ist orderTransition Probability Matrix (TPM) (i.e. first order Markov Model) is evaluated. In this, each state is composed of only single page as depicted in Table 3. Then, second-order Markov model is evaluated. In this each state will be composed of two web pages and this is decided by the entries in the first-order TPM as shown in Table 4 and so on.

This transition probability matrix can be now use to predict the next click for the given session. For example, consider a user's navigation sequence as **D—>E—>?** To predict the next page after D and E, firstly the state {D, E} is identified in the second-order TPM and then the page with highest probability is selected. Here, B has the highest probability among rest of the pages as seen from Table 4.

Therefore, **D—>E —>B** is obtained.

Similarity Calculation: For instance, consider two pages A and B with their respective page URLs as shown in Table 5. Thus, the similarity of the two pages, calculated by the method above explained, comes out to be (4+2+1) / (4+3+2+1) = 0.7. It indicates that the pages are somewhere similar but not exactly same.

Advantages and Limitations of PSPR: The main advantage of this method is that it improves the prediction of web page access by analyzing web users' navigational patterns. It can be applied to any web site's navigational graph for improving browsing orders. But, this method fails to predict directly one more step ahead.

C. SIMRANK: PAGE RANK APPROACH BASED ON SIMILARITY MEASURE

Shaojie Qiao et. al [10] proposed a better and promising approach to rank the web pages in the query result liston the basis of the similarity measure from the vector space model named as SimRank. This method computes the similarity of pages and applies it to partition the whole web database into various web social networks (WSNs). This method utilizes the concept of social annotations [11] named as SimRank. The web annotators associate some set of textual content with every web page so as to provide a prior knowledge regarding the web page to the web user without reading the internal contents of that page. In other words, they provide a brief overview about the web page and thus make the user's navigation fruitful. These set of textual contents are known as annotations. The annotations are parsed contents holding the important keywords of a web page. Thus, this method considers the similarity measure from vector space model so as to determine the rank of pages. It also improves the traditional PageRank algorithm by considering the relevance of a web page to a given query.

When a web user fires a query using suitable keywords to fetch the required web content, some web pages are retrieved that possess both the relevant and irrelevant pages as per the query. This method assigns a relevance score to each retrieved web page based on the similarity between the query keywords and the annotations. In simpler terms, it matches the contents of the query with the annotations of every web page and accordingly determines their importance against the query.

SimRank works in the following manner:

- Firstly, it computes similarity among the web pages of the complete web database.
- Then, it uses the similarity measure as the distance between the pages and apply K-means algorithm to form clusters with pages holding similar contents, and
- Finally, it computes the similarity with respect to the query and assigns a relevance score to each web page. But, this method has issue that its efficiency gets affected by the capabilities of the web crawler being utilized.

The term frequency of a term t_i in the page d_j is calculated byusing (3),

$$tf_{ij} = \frac{f_{ij}}{\max\{f_1, f_2, \ldots f_{|V|j}\}} \qquad (3)$$

where f_{ij} denotes the frequency of the term t_i in the page d_j and |V| is the vocabulary size.

The inverse document frequency of term t_i is given by using (4),

$$idf_i = \log\left(\frac{N}{df_i}\right) \qquad (4)$$

where N is the total number of web pages in the web database, df_i denotes the number of web pages in which the term t_i appears atleast once.

Now, the overall term weight is computed as in (5):

$$w_{ij} = \left\{0.5 + \frac{0.5 \times f_{ij}}{\max\{f_1, f_2, \ldots f_{|V|j}\}}\right\} \times \log\frac{N+1}{df_i} \qquad (5)$$

The similarity measure of a query $Q = \{t_1, t_2 \ldots t_n\}$ and a page p_j denoted as $p_j = \{w_{1j}, w_{2j}, \ldots, w_{nj}\}$ where n is the number of terms in the query. So, similarity between two pages p_a and p_b is computed by using (6):

$$sim(p_a, p_b) = \frac{\sum_{i=1}^{n} w_{ipa} \times w_{ipb}}{\sum_{i=1}^{n} w_{ipa}^2 + \sum_{i=1}^{n} w_{ipb}^2 - \sum_{i=1}^{n} w_{ipa} \times w_{ipb}} \qquad (6)$$

Example Illustrating Working of SimRank.

To illustrate the working of SimRank, let us consider two papers with their contents as shown below,

P1 = The project objectives are laid down as per the required project.
P2 = Organisation signs a project yesterday.

And the query entered is {project}.Based on the equations (3), (4), (5) and (6), we obtain:

$$tfp_1 = \frac{2}{11} = 0.181 \qquad tfp_2 = \frac{1}{5} = 0.2$$

$$idf_{p1} = \frac{4}{2} = 2 \qquad idf_{p2} = \frac{4}{2} = 2$$

$$w_{p1,p2} = (0.5 + 0.5 \times 0.181) \times 2 = 1.181$$

$$w_{p2,p1} = (0.5 + 0.5 \times 0.2) \times 2 = 1.2$$

$$sim(p_1, p_2) = \frac{1.181 \times 1.2}{1.181^2 + 1.2^2 - (1.181 \times 1.2)} = 0.99$$

Thus, it shows that the contents of both the papers are relevant as per the query entered. Hence, SimRank judge with better accuracy about the papers against the content interested to the user.

Advantages and Limitations of SimRank: The main advantage of this method is that it uses similarity measures to effectively cluster and score the publications. This method uses k-means clustering approach to subdivide a web database into various WSNs as well as clean up the irrelevant and unrelated pages that can help in reducing the cost of computation. But, this method has issue that its efficiency gets affected by the capabilities of the web crawler being utilized.

D. *PAGE RANKING USING SOCIAL ANNOTATION BASED ON LANGUAGE MODEL*

Kunmei Wen et. al. [12] proposed an extension to SimRank named *optimizing the results with social annotations based on a language model*. This method uses social annotations to re-rank search results. This method uses the combination of two ranking strategies (a) query-annotation similarity, and (b) query-document similarity in order to optimize retrieval ranking method.

This method works in the following phases:

- To build the statistical language model of social annotation.
- Calculated the similarity among query and annotation using the language model.
- Initial results of a search engine are re-ranked on the basis of combined score of both the similarity measure.

Statistical language model: In this method, the input parameters considered for constructing a language model are as:

- Set of K initial search results denoted as D= $\{(R_1, A_1)\ldots(R_k, A_k)\}$ produced by a search engine where R_k denotes the page and A_k denotes a set of annotations against a specific R_k.
- Set of social annotations in the top K initial search results (also refer as a temporary corpus) denoted as $V_A = \{W_j \mid j = 1 \ldots L\}$ where L denotes the size and W_j is a social annotation.
- Set of the social annotations of a specific page denoted as $A_i = \{a_i \in V \mid i = 1, \ldots, n\}$

The steps involved in the language model construction include:

- Identify the annotations associated with the web pages and initialize the set A_k accordingly.

- Derive temporary corpus (or the collection of all the social annotations) from the K initial search results.
- Calculate the probability of a term denoted by w_i in the set of annotations A_i for a specific web page using the formula as shown in (7),

$$P(w_j|A_j) = \frac{C(w_j, A_j) + 1}{\sum_w (w_, A_j) + L} \qquad (7)$$

- Thus, it results in the K language models of the annotations for top K initial results.

Query-annotation similarity:
User enters the query in the form of keywords, therefore a query can be denoted as Q = $\{q_1, q_2 \dots\dots q_m\}$ where q_i refers to the keywords or corpus. The probability of the existence or generation of a specific query Q in A_i 's language model is represented as P(Q|Ai). This is referred as probability of query generating. The similarity computation between query and annotations involves the following steps:

- Firstly, the probability of terms appearing in specific annotation is derived from the language model of social annotation.
- A weight is assigned on the similarity measure between query and social annotation and the results are stored.
- The frequency count of a term w in the given query Q is represented as C(w,Q) is taken into account to contribute in similarity score.
- Similarity weight between query and annotation is calculated by using (8),

$$P(Q|A_i) = \prod\nolimits_{w \in Q} P(w|A_i)^{C(w,Q)} \qquad (8)$$

Final Rank Score: This method finally calculates the rank score of paper by integrating the query-annotation similarity denoted by P(Q|R) and query-document similarity denoted by P(Q|A). The combined weighted rank score is calculated by using (9),

$$Score_i = \alpha \times P(Q|R_i) + \beta \times P(Q|A_i) \qquad (9)$$

where α and β are weights determined experimentally and satisfy $\alpha + \beta = 1$ equation.

Advantages and Limitations: This method uses the concept of annotations which are used as a brief summary for a publication. By using this approach the results are optimized and the newly formed search list is more accurate. But sometimes these annotations contain incomplete and unrelated terms. Such annotations are considered as sparse in nature.

III. COMPARATIVE STUDY

By thorough study of web mining concepts and analysis of important paper ranking algorithms, it is concluded that every algorithm is unique and relevant in itself. They persist and are currently in use because of their strengths and individuality in performing ranking of papers. Also, due to some limitations they do require further improvements for serving with more accuracy and relevance. A detailed comparison of ranking algorithms studied is listed in Table 6 on the basis of various measures such as main techniques used, methodology, input parameters, relevancy, quality of results, importance and limitations.

IV. CONCLUSION

Generally, people rely on digital library search systems to explore the web for scientific information. In such a scenario, many advanced web searching and mining techniques have been employed in order to find only relevant information. Page ranking algorithms play a significant role in ranking scientific information, so that the user can retrieve the most relevant information in the top of the result list. Some algorithms rely on the link structure of the publications (web structure mining), whereas others consider the contents in the publications (web content mining), while some are based on the combination of both. Therefore based upon the technique used, the ranking algorithms present a different order of resultant publications. As part of future work, an efficient page ranking algorithm in terms of time response, accuracy, and importance of the results; and relevancy of results should be developed and implemented so that the quality of web search results can be improved.

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] 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.

[4] 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

[5] 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

[6] 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.

[7] A.K. Sharma, N. Dhuan and G.Kumar, "A Novel Page Ranking Method based on Link-Visits of Web Pages", Int. J. of Recent Trends in Engineering and Technology, vol. 4, no. 1, pp. 58-63, Nov 2010.

Table 6. COMPARISON OF RANKING ALGORITHMS

Algorithms → / Measures	**PageRank**	**PSPR**	**SimRank**	**Social Annotation based on language model**
Main Technique used	Web structure mining	Web Usage Mining, Web structure mining	Web Content Mining	Web content mining
Description	Link analysis based algorithm that computes importance on the basis of backlinks.	The search result list is ranked based on Markov model output and frequency of transition and similarity of papers.	Papers are ranked according to the content similarity rather than the link structure of the pages.	Result are ranked based on the weighted scored determined by calculating similarity score between query and annotation as well as query and document
Input Parameters	Backlinks	Web sessions (Sequence of pages accessed).	Papers and query contents.	Initial search result list, set of tags and papers.
Complexity	O(logN) where N denotes the web pages	O(N) , where N denotes the number of pages (or states) .	$O(N^2)$ where N denotes number of papers.	O(K*L) where K denotes size of the K initial result list and L denotes size of the temporary corpus.
Relevancy	Less	More relevant than traditional PageRank Algorithm.	Results obtained are relevant than the traditional PageRank and other extensions of PageRank.	More relevant results than in SimRank approach.
Quality of results	Medium	Markov models are highly vulnerable to the data set being used.	Increased efficiency and accuracy in ranking of pages in result list.	This method highly optimizes the initial search results that use only query-document similarity.
Importance	Simplest method that analyses whole link structure of graph at once to compute relevance.	Improves the prediction of web page access & can be applied to any web site's navigational graph for improving browsing orders.	Effectively analyze pages or documents with little contextual information.	It optimizes the ranking of initial results by integrating query-annotation similarity with query-document similarity.
Limitations	Results obtained at the time of indexing and not at the query time.	It fails to predict directly one more step ahead.	Its efficiency gets affected by the capabilities of the web crawler being utilized.	In some web pages the annotations may be sparse and incomplete; hence it creates a gap between annotations and queries.

[8] P. Thwe, "Proposed Approach For Web Page Access Prediction Using Popularity And Similarity Based Page Rank Algorithm", international journal of scientific & technology research, vol. 2, no. 3, march 2013.

[9] M. Deshpande and G. Karypis, "Selective markov models for predicting web page accesses", ACM Trans. Internet Technol, vol. 4, pp. 163-184, May 2004.

[10] S. Qiaot, T. Li, H. Li, Y. Zhu, J. Pengt and J. Qiu, "SimRank: A Page Rank Approach based on Similarity Measure", IEEE, 2010.

[11] V.T. Nguyen, "Using social annotation and web log to enhance search engine", IJCSI International Journal of Computer Science Issues, vol. 6, no. 2, 2009.

[12] K. Wen, R. Li, J. Xia and X.Gu, "Optimizing ranking method using social annotations based on language model", Springer Science+Business Media B.V. 2011, Jan 2012.

[13] N. Tyagi, S. Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", International Journal of Soft Computing and Engineering (IJSCE), vol. 2, no. 3, July 2012.the first word in a paper title, except for proper nouns and element symbols.