Comparative Analysis of Page Ranking Algorithms in Digital Libraries

Suruchi Nehra
M. Tech Student
CSE & IT Department
The NorthCap University
Gurgaon,India

Deepti Gaur
Associate Professor
CSE&IT Department
The NorthCap University
Gurgaon,India

ABSTRACT

Page ranking algorithms are important facet of ranking the articles in online digital libraries. Researchers utilize the digital libraries for their research and to find popular, recent and relevant articles in their domain. Ranking plays crucial role in searching, as there are millions of articles present in academic digital libraries, there is a need to order them so that users can find propitious articles efficiently. This paper presents a study and a comparative review of various ranking algorithms in online digital libraries under different web mining techniques based on their scope, performance, advantages and challenges. The paper also shows that how some of the drawbacks of certain algorithms are met by other proposed algorithms. This comparative analysis helps in further improvements in the related field.

Keywords

Digital Libraries, Page Ranking, Search Engine, Web Usage Mining, Web Content Mining, Web Structure Mining.

1. INTRODUCTION

In the evolving world internet has become one of the essential tool for accessing and sharing the varied information, as a result information is swiftly increasing with the growth of World Wide Web. In 2013 alone, the web has grown by more than one third [1] and number of users estimated are 2,802 millions [2]. Due to increase in information sources and requirements of users there is a need to manage and process the information in such a way that the user can find desired results efficiently. Even though there is advancement in a general purpose search engines; there exist a situation when users or researches find irrelevant results for a given query. For example if a user search for research papers, journals or books on certain topic, a search engine returns a list containing blogs, articles, news etc. To overcome this problem digital libraries or digital repositories have been developed so that desired results are made available to users. Digital library is an organized and focused collection of digital content with the methods of access and retrieval and for the selection, creation, organization, maintenance and sharing of collection [3] .For example: Science Direct, IEEE Xplore, Springer, ACM Digital Library are popular ones. The exponential growth in the quantity and diversity of digital library's content signifies both challenges and opportunities. One of the challenges is to provide users with most relevant, valuable, useful and best information in less time .As a result most advance search engine technologies are employed in digital libraries .The paper aims to compare and classify some of the ubiquitous page ranking algorithms in online digital libraries. The paper is structured as follows: Section 2 presents some of the page ranking algorithms under different categories with their advantages and limitations. In Section 3 comparison study based on different parameters is carried out. Finally, Section 4 outlines conclusion and future work.

2. PAGE RANKING ALGORITHMS IN DIGITAL LIBRARIES

Page ranking algorithms in digital libraries are categorized on the basis of type of web mining being used. Some algorithms use only link structure of papers (web structure mining), whereas others use only content of papers (web content mining), while some rely on server log files (web usage mining) and some use combination of these mining techniques.Some of the algorithms under different categories have been discussed as follows.

2.1 Algorithms under Structure Mining

2.1.1 Citation count ranking algorithm

Joeran Beel et al [4] proposed Citation count ranking algorithm in 2009, which is one of the simplest methods for ranking the publications. The algorithm is based on the citation graph without any other parameter included in it. Citation graph can be viewed as directed graph G(V,E) where the vertices V indicate the publications and the edges E indicate the citations between the publications, weight of each $v_i \in V$ defined as follows

$$W(v_i)= \sum_{j=1}^{j=|v|} W(e_{ji}), \forall_i = (1, |v|) \text{or} \overline{1, |v|} \quad (1)$$

Where W (e_{ji}) is weight of each edge, and it is define as follows

$$W(e_{ji}) = \begin{cases} 1 \text{ if } j \rightarrow i, \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

According to the algorithm, publications having more citation count are considered more important than others and they are given a high rank. Citation count of paper indicates the number of readers who refer to it; hence it is one of the parameter in support of goodness and popularity of an article. A publication having more citation is likely to acquire top position in result list of user's query.

The advantage of the algorithm is that it is simple to implement and it gives top of most cited publications. However it overlook the importance of the citing papers and handle all citations in the same way .It also does not consider the time factor due to which recent papers are ranked lower as compared to older papers as they get less time to acquire more citation count. It also increases the Matthew Effect [5] which means that highly cited papers are displayed first by the search engines therefore they get more citations from readers which in turn make them to be displayed first again.

2.1.2 Page rank algorithm

Surgey Brin and Larry Page [6], developed a link-based ranking algorithm in 1998, named Page Rank. According to the algorithm if a publication has some important incoming link to it then its outgoing links to other publication also become important, which can be defined as

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

Where d is a dampening factor that is generally set to 0.85, d represents the probability that user will follow new link rather than following a link on the current page. B (u) represents the set of pages that refer to u, PR (u) and PR (v) are rank scores of page u and v, N_v indicates the number of outgoing links of page v. As already discussed, the major drawback of citation count algorithm is that it treats all the citation equally. Page rank overcomes this limitation by assigning high weights to the publications that are cited by important papers. In this way, the algorithm helps in recognizing those articles that contain important work for the researchers.

The main advantage of Page Rank algorithm is that it ranks the publications based on the importance of their citations and gives high rank to important articles that would not have been found by using citation count method but the algorithm behaves differently in case of incomplete citation graph containing missing citations. For example, a publication that has only one reference in the citation graph, will forward all its weight to the single reference and if the publication has high page rank then the only reference will gather a lot of artificial weight. It also gives more weightage to older articles as compared to recent ones, and does not take into account other parameters like author, venue, time etc. Moreover the method gets affected from outgoing links [7],which means if a paper P is cited by many papers with high ranks but have a large number of outgoing links, then paper P's rank get decrease.

2.1.3 Time-dependent Link-based Ranking algorithm

Martin Rajman and Jean-Yves Le Meur [8] in 2009 proposed a new ranking method named as time–dependent link-based ranking .The algorithm combines the idea of time dependent citation with page rank algorithm to rank recent and important publications higher. The ranking method weighs each publication inversely proportional to its age and gives more value to the citations of recent publications; hence the initial probability of choosing the i^{th} paper in a citation graph is given by

$$p_i = e^{-w(t - t_i)} \qquad (4)$$

Where t denotes the present time, t_i is the publication date for i^{th} publication and w is the time decay parameter. This initial probability is added to page rank equation (3) as follows

$$\mathrm{PR}(u,t) = \sum_{x=1}^{n} \left(\frac{(1-d)}{n} \times P_x(t) \right) + d \sum_{v \in B(u)} \left(\frac{\mathrm{PR}(v)}{N_v} \times (t) \right) \qquad (5)$$

Where $P_x(t)$ is the initial probability of selecting the x^{th} node in the citation graph and n is the total number nodes in the graph. The algorithm overcomes the two main problem of citation count method as discussed above by utilizing both time of citation and importance of the citing paper.

The main advantage of the algorithm is that it gives high rank to the recently cited and important publications but the method overestimates the weight of recent publications that are part of a cycle therefore it is not suitable for data set that allow cycles.

2.1.4 Focused page rank algorithm

Mikalai Krapivin and Maurizio Marchese [9] proposed Focused page rank (FPR) algorithm in 2008.This method suffers less from outgoing links problem as compared to the page rank. The proposed algorithm is a tradeoff between Page Rank and Citation Count Method. Page rank is based on random surfer model which means that page rank of particular node depends on the probability to arrive at this node by randomly traversing the graph. At each step link to be follow is selected randomly, but the focused surfer becomes focused by selecting the path which is more preferable for him. Which can be define as

$$P_i = (1\text{-}d).\sum_{i \neq j}^{j \in D} P_j . S(j|i) + \frac{d}{N} \qquad (6)$$

Where S (j|i) is the probability to follow the reference i being at the page j. S is an arbitrary function. Simplest version of it is define as follows

$$S(j|i) = \frac{c(i)}{\sum_{K \in D} C(K)} \qquad (7)$$

Where C (m) is paper m citations count and D is the set of all references in paper C (j).

The main advantage of the algorithm is that it is simple to implement and it involves benefits of both quantity of citations and the quality of citations. The method suffers less from the effect of outbound links. However, it does not take into account other parameters like author, venue, time etc and gives more weight age to older articles as compared to recent ones.

2.1.5 link-based Ranking with External citations

Martin Rajman and Jean-Yves Le Meur [8] in 2009 proposed a new ranking method based on external citation, named as external citation ranking. The algorithm is made for the data sets having incomplete citation graph i.e graph containing missing citations, it assume that there is new node called "external authority" that collect weight from all the nodes in graph proportionally to the missing citations and feedback some amount of its weight into network. As already discussed, when page rank algorithm is applied over the network containing the missing references, it distributes the weight to only those references that are present in the paper without considering the missing references, so these present references gain much more weight than usual. External citation ranking overcomes this drawback of page rank by assuring the accurate spread of weights through the network. It also corrects the main problem of Citation count method by considering the importance of the citing paper.

The algorithm is better than both the methods of ranking i.e. citation count and page rank as it corrects the major shortcoming of these two methods. But it does not take into account other parameters like author, venue, time etc and ranks older articles higher than recent ones.

2.2 Algorithms under Content Mining

2.2.1 Simrank: Algorithm based on similarity

Shaojie Qiao [10] proposed a better ranking algorithm named as SimRank in 2010.The algorithm ranks the publications on the basis of similarity factor based on vector space model(IR model)[11],[12] and also uses the similarity parameter to segregate the entire web database into different web social networks (WSN) .The algorithm works on the basis of assigning relevancy score to each of the retrieved pages from the user query by comparing the content of query with title and body of every page. The algorithm computes Term Frequency of term t_i in the page d_j by using

$$tf_{ij}=\frac{f_{ij}}{\max\{f_1\ ,f_2\ ,\ldots\ldots\ldots f_{|v|j}\}} \quad (8)$$

Where f_{ij} denotes the frequency of the term in the page d_j and |V| is the vocabulary Size. The Inverse document frequency of term is defined as

$$df_{i\,=}\ \log(\frac{N}{df_i}) \quad (9)$$

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 appears at least once and overall term weight is calculated as

$$w_{ij} =\left\{0.5+\frac{0.5\times f_{ij}}{\max\{f_1,f_2,\ldots\ldots f_{|v|j}\}}\right\}\times\log\frac{N+1}{df_i} \quad (10)$$

Now, similarity between two pages p_a and p_b is computed by

$$sim(p_a,p_b)=\frac{\sum_{i=1}^{n} w_{ipa}\ \times w_{ipa}}{\sum_{i=1}^{n} w_{ipa}^2+\sum_{i=1}^{n} w_{ipb}^2-\sum_{i=1}^{n} w_{ipa}\ \times w_{ipa}} \quad (11)$$

The Algorithm works in following steps:

i. In first step, it calculates similarity among pages of the whole web database by using equation (11)

ii. Then, it employs similarity value as distance between the papers and use k-means algorithm to make the clusters of pages having the similar content.

iii. It calculates the similarity on the basis of query and assign relevancy score to each paper.

The main advantage of this algorithm is that it uses similarity factor in K-means clustering to segregate a web database into different WSNs. It also removes irrelevant and unrelated pages in order to reduce the cost of computation. But main problem with this method is that its efficiency gets influence by the capabilities of the web crawler being exploited.

2.2.2 Page ranking using social annotation based on language model.

Kunmei Wen et. al.[13] proposed a ranking algorithm which is an extension to simrank in 2012, named as page ranking using social annotation based on language model. The algorithm optimizes the result by using the concept of social annotation [14]. Annotations are used for re-ranking the initial search results. The method uses two techniques known as query-annotation similarity and query-document similarity. The algorithm first builds language model of social annotation .Then similarity between query and annotation is computed with the help of language model. In the end, initial search results are re-ranked on the basis of collective score of both the similarity techniques.

Statistical language model: It is use in information retrieval(IR) model. The model uses following input parameters: a) Set of K original search results of search engine, define as D={(R_1,A_1)….($R_{K,}$,A_k)}where R_k represents page and A_k represents is a set of annotations in a particular page R_k. b) Set of social annotations associated with top K initial search results define as V_A ={W_j| j=1…L} Where L represents size and W_j is social annotations in the top K initial search results.c) Set of social annotations of a particular page define as A_i= {a_i ∈V| i = 1, . . . , n} and Steps for the construction of language model are:

i. Initialize the set A_k with annotations related to web pages.

ii. Obtain temporary corpus from the k initial results.

iii. Compute the probability of a term indicated by w_i in the set of annotations A_i for a particular page by using following equation

$$P(w_j|A_j)=\frac{C(w_j\ ,A_j)+1}{\sum_w(w_j,A_j)+L} \quad (12)$$

iv. Output the k language model of the annotations for top k initial results.

Query-annotation similarity: Query is represented as Q ={q_1,q_2,q_3….q_m} where q_i refer to keywords, the probability of existence of particular query Q in A_i ‘s language model is denoted as P(Q|A_i), calculation of similarity between query and annotation involve following steps:

i. Probability of term appearing in particular social annotation is derived from language model.

ii. Weights are assigned on the basis of similarity measure between the query and social annotation.

iii. Frequency of the term w in the given query Q i.e C(w,Q). is calculated for the computation of similarity score

iv. Similarity weight between query and annotation is computed by

$$P\,(Q|A_i)=\prod_{w\in Q} P(w|A_i)^{C(w,Q)}$$

(13) Final Rank Score: Final rank score is computed by

$$Score_i=\alpha\times P(Q|R_i)+\beta\times P(Q|A_i) \quad (14)$$

Where P (Q|R) is query-annotation similarity, P (Q|A) query-document similarity and α and β are weights determined experimentally and satisfy the equation α+ β=1.

The main advantage of the method is that it uses the idea of annotation which is a brief information about the publication. It also gives more optimized and accurate result. However Annotation may contain incomplete and unrelated terms therefore it can decrease the performance of the algorithm.

2.3 Algorithms under More Than one Mining Technique

2.3.1 Futurerank : Ranking Articles by Predicting their Future Page Rank

Hassan Sayyadi, Lise Getoor [15], proposed new ranking algorithm named as future rank in 2009.The algorithm considers the dynamic nature of citation graph, as publications acquire new citations every day. According to the algorithm two factors play important role in ranking of publications, that are popularity and usefulness of an article. Popularity of an article can be determined by number of current citations at the time of query and usefulness of an article can be determined by expected future citations. The method define a new parameter called as future rank, it is the expected future rank of a paper based on the citation that paper is going to get in future. The algorithm uses authorship network, publication time of an article and citation network to predict future citation. Page rank algorithm is applied on citation network and HITS algorithm [16] is applied on authorship network. Networks can be represented by adjacency matrices as follows

$$M_{i,j}^{C}=\begin{cases}1 \text{ if } p_i \text{ cites } p_j;\\ 0 \text{ otherwise;}\end{cases}$$

(15) For any paper p_i which does not cite any article in the dataset $M_{i,j}^{C}$ = 1 is define for all j. The matrix M^A which is the |P| × |A| authorship matrix is defined as:

$$M_{i,j}^{A} = \begin{cases} 1 \text{ if } a_i \text{ is author of } p_j; \\ \quad 0 \text{ otherwise;} \end{cases} \qquad (16)$$

The algorithm works on both the networks by passing information back and forth between the networks. It run one step of page rank and one step of HITS and combine their results thus, it forms an iterative algorithm by repeating the steps until convergence is obtain .Vectors of page score and author rank is denoted by R^P and R^A respectively. Hub score of authors is calculated by $R^A = M^A * R^P$ and rank of papers is computed by following equation

$$R^P = \alpha * M^C * R^C + \beta * M^A * R^A + \gamma * R^{Time} + (1 - \alpha - \beta - \gamma) * [1/n]. \qquad (17)$$

Where $M^c * R^c$ is page rank score in citation network, $M^{A^T} * R^A$ is authority score in authorship network ,α β γ are parameters which weights the citation , author and publication time respectively. R^{Time} is a "personalized page rank vector" its default value is 1/n for all nodes, calculated by (18) and values in vector are pre-computed based on the current publication time of papers i.e. $T_{current}$

$$R_i^{Time} = e^{-p*(T_{current} - T_i)} \qquad (18)$$

Initial values of R_i^P and R_i^A is $\frac{1}{|P|}$ and $\frac{1}{|A|}$ and $\alpha + \beta + \gamma + (1 - \alpha - \beta - \gamma)$ is equal to one.

The main advantage of the future rank algorithm is that it combines citations, authors and publication time in an effective way for predicting article's future rank. It achieves better improvement over other recently proposed algorithm as it has faster model convergence and high correlation score. But it does not guarantee robustness on different values of parameters α, β, and γ in different dataset and does not take into an account the importance of publication venues such as conferences and journals.

2.3.2 Ranking articles by using citations, authors, journals and time information

Yujing Wang,Ming Zeng [17] proposed a ranking algorithm which is an extension to future rank in 2013.The algorithm overcomes the shortcoming of future rank by exploiting journal information along with citations, authors and time Information. The method constructs heterogeneous network which contains three sub-networks (citation network, paper-author network, and paper-journal network). Citation network contains only papers nodes and citation edges, paper-author network contains paper nodes and author nodes with authorship edges between them and it forms a bipartite graph. Paper journal network contains paper nodes and journal nodes and undirected edge between a paper and a journal/conference. The algorithm first assigns value $\frac{1}{N_P}$ to all the authority scores of papers where Np is the number of papers in the collection. Then it calculates Hub score of authors, journals/conferences, papers by paper-author network, paper-journal network and citation network respectively by following equation

$$H(x) = \frac{\sum_{P_j \in Neighbor(x)} S(P_j)}{|Neighbor(x)|} \qquad (19)$$

Where x denotes the parameter whose hub score is to be calculated, parameter can be author ,journal and paper, H(x) is the hub score of a parameter, it can be $H(A_i)$for author A_i,, $H(J_i)$ for the journal/conference J_i , $H(P_i)$ for paper P_i, $S(P_j)$ is the authority score of paper P_j, Neighbor (x_i) is the collection of papers which correspond to parameter x, and |Neighbor(x)|is the number of papers in the collection. The algorithm then update the authority scores of each paper by using Page Rank score contributed from citation, authors score, journals/conferences score and time-aware score calculated using publication date by using following equation.

$$S(P_i) = \alpha \cdot PageRank(P_i) + \beta \cdot Author(P_i) + \gamma \cdot Journal(P_i) + \delta \cdot Citation(P_i) + \theta \cdot P_i^{Time} + (1 - \alpha - \beta - \gamma - \delta - \theta) \cdot / N_p \qquad (20)$$

Where $S(P_i)$ is updated authority score of paper p_i and α, β, γ, δ and θ are constant parameters which ranges in (0, 1) . Page Rank (P_i) is the page rank score of paper p_i ,calculated by citation network. Author (P_i) is authority score of paper $P_{i,}$ calculated using paper-author network as follow

$$Author(P_i) = \frac{1}{Z(A)} \sum_{A_j \in Neighbor(P_i)} H(A_j) \qquad (21)$$

Where Neighbor(P_i) is the author list associated with paper P_i, and H (A_j) is the hub score of author A_j. Z(A) is a normalized value. Journal(P_i) is authority score of paper P_i generated from corresponding journal/conference and Citation(P_i) is authority score of paper P_i gathered from hub papers. P_i^{Time} is a time aware value for paper p_i define by

$$P_i^{Time} = e^{-p*(Tcurrent - Ti)} \qquad (22)$$

Where T_i is the publication time of paper Pi, $T_{current} - T_i$ is the duration in years since the paper Pi was published. p is a constant value, which is set to 0.62 and (1−α−β −γ −δ−θ)*1 /Np is the probability of random jump. The whole procedure is repeated until convergence is obtained.

The main advantage of the algorithm is that it promotes the recent articles by giving higher scores to them. It takes combination of different types of information and time aware weights for better ranking of articles. However, it requires compact graph to obtain accurate result.

2.3.3 Popularity and similarity based page rank algorithm.

Phyu Thwe [18] proposed a ranking algorithm in 2013 for web page access prediction, named as Popularity and Similarity Based Page Rank Algorithm. The method is refinement over the prediction of pages access by user and uses the web server log files for analyzing the user's browsing pattern in order to predict the user's next click. It ranks the result of search engine on the basis of three factors a) popularity b) similarity among pages and c) user's browsing pattern. The algorithm works in two steps:

i) Markov model construction: This step uses Markov model for prediction of web page access, input of the model is pages taken in the order of browsing by user and output is a model that predicts the user next access. The model assumes P be a set of web pages in a web site and W be a user session of a website . P can be written as P= {p_1, p_2...p_n} .Therefore the probability of accessing the next page p by the user is indicated as P = (p_i|W), it is base on the assumption that i number of pages has already been accessed by the user. It can be infer, that prediction of next page to be accessed does not depend on all the pages present in web session but depends on small number of k pages. Where k is order of markov model. So page p_i+1 can be accessed by using following equation.

$$p_{i+1} = \text{argmax}_{p \in P} \{P(P_{i+1} = p | p_i | p_{i+1}, \dots \dots . p_{i-(k-1)})\} \qquad (23)$$

ii) Similarity Calculation: In this step popularity, similarity and transition among pages are determined to compute the importance of the pages. Similarity is computed based on the contents of the page URL by using the following step: a) first, choose the URLs of the two pages in order to compute similarity between them. b) The URLs are sorted in a string array separated by a special character '/' and their length is computed. c) Weights are given to each array beginning from the longest array to the smallest one. d) The matching substrings are recognized and their equivalent weights are summed up and the sum is divided by the total weight to provide the similarity value between the two. Similarity value can occur between 0.0 and 1.0, value 1 denotes that the two pages are identical, value 0 indicate pages are entirely different.

One of the main advantages of this algorithm is that it makes prediction method better by inspecting the user's browsing orders and it can work on any website's navigational graph But the method fails in case of predicting one more step ahead.

3. COMPARISON STUDY

Comparison of page ranking algorithms studied so far in different categories is done on the basis of input parameter, importance and limitations. Comparison of algorithms lying under structure mining is done in Table 1 and comparison of algorithms lying under content mining is done in Table 2. Comparison of algorithms utilizing combination of different web mining techniques is done in Table 3.

4. CONCLUSION

The paper presents different page ranking algorithms in digital libraries, each of these algorithms has its own advantages, limitations, mining techniques, and performance. These Algorithms provide different result set based on the mining technique being used. Some of the algorithms are proposed in response to counteract the drawbacks of earlier proposed algorithms, while for some algorithms various extensions and improvements might be thought. As a further research issue, there can be several options; one of them would be to develop more efficient algorithms that are likely to satisfy researcher's needs and desires by fruitfully combining parameters like time, author, venue of publication, citations and mining techniques. Another can be to experiment with the available algorithms in order to overcome their limitations and to improve them in terms of their response time, accuracy and performance. As page ranking algorithms are use online and should be fast and accurate, therefore soft computing approaches can be applied for near optimal solutions.

5. REFERENCES

[1] Total number of Websites Internet Live Stats, http://www.internetlivestats.com/total-number-of-websites/

[2] Internet World Stats – Usage and Population Statistics ,www.internetworldstats.com

[3] Kumar,V.Vijay,P.Rama MohanRao,Digitization of Library Resources and the Formation of Digital Libraries: A Practical Approach, International Journal& Magazine of Engineering, Technology, Management and Research, Volume No: 1, Issue No: 12 ,2014,pp69-71.

[4] Beel, Jöran, Bela Gipp, Google Scholar's ranking algorithm:the impact of citation counts (an empirical study),Third International Conference on Research Challenges in Information Science, 2009, pp. 439–446.

[5] Merton, Robert K,The Matthew effect in science, Science Vol.159. No.3810,1968 ,pp 56-63

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

[7] Sobek M, The effect of outbound links, Internet paper, http://pr.efactory.de/e-outbound- links.shtml.

[8] L. Marian, M. Rajman, Ranking Scientific Publications Based on Their Citation Graph,Master Thesis, CERNTHESIS,2009.

[9] Krapivin, Mikalai, and Maurizio Marchese,Focused page rank in scientific papers ranking , Digital Libraries: Universal and Ubiquitous Access to Information,Springer Berlin Heidelberg, 2008, pp. 144-153.

[10] S. Qiaot, T. Li, H. Li, Y. Zhu, J. Pengt , J. Qiu, SimRank: A Page Rank Approach based on Similarity Measure, 2010 International Conference on Intelligent Systems and Knowledge Engineering (ISKE) ,IEEE 2010, pp. 390-395.

[11] D. A. Grossman and O. Frieder, Information Retrieval: Algorithms and Heuristics. Springer, Secaucus, NJ, USA, 2004.

[12] G. Salton and M. McGill, An Introduction to Modern Information Retrieval, McGraw-Hill, New York, NY, 1983

[13] K. Wen, R. Li, J. Xia and X.Gu, Optimizing ranking method using social annotations based on language model, Artificial Intelligence Review, 2014,41(1) ,pp.81-96

[14] V. T. NGUYEN, Using social annotation and web log to enhance search engine, International Journal of Computer Science Issues, IJCSI, Volume 6, Issue 2, 2009, pp1-6.

[15] Sayyadi, Hassan and Lise Getoor ,Futurerank: Ranking scientific articles by predicting their future pagerank,Proceedings of the Ninth SIAM International Conference on Data Mining (SDM'09),2009,pp. 533–544.

[16] Kleinberg J,Authorative Sources in a Hyperlinked Environment, Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval(46) ,no. 5,1999,pp 604-632.

[17] Yujing Wang, Yunhai Tong, Ming Zeng, Ranking Scientific Articles by Exploiting Citations, Authors, Journals, and Time Information, Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence,2013,pp.933-939.

[18] 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. 3,2013,pp. 240-246

6. APPENDIX

Table 1.Comparison of algorithms under structure mining

Algorithms → Measures ↓	**Citation count**	**Page rank**	**Time-dependent Link-based Ranking algorithm**	**Focused Page rank**	**Link-based ranking with external citations**
Description	The importance of the paper is based upon the number of citations to it.	It is the Link analysis based algorithm that computes importance on the basis of backlinks.	The algorithm combines the idea of time dependent citation with page rank algorithm to rank recent and important publications higher.	It is a trade off between Page Rank and Citation Count Method and uses the focused surfer model and selects the path which is more preferable for him.	It assume the presence of external authority(a new node) that collect weight from all the nodes in graph proportionally with the missing citations and feedback some amount of its weight into network.
Input parameters	Citation count	Backlinks	Citation count, Backlinks	Citation count, Backlinks	Citation count, Backlinks
Importance	It gives top of Most Cited Publications.	It weights the publication based on the importance of their citation.	it gives high rank to the recently cited and important publications.	It suffer less from the effect of outbound links.	It is better than both the methods of ranking i.e citation count and page rank.
limitation	Ignore the importance of citing paper and treats all citation equally	give more weightage to older articles as compared to recent ones	It overestimate the weight of recent publications that are part of a cycle	give more weightage to older articles as compared to recent ones	Give more weightage to older articles as compared to recent ones

Table 2. Comparison of algorithms under content mining

Algorithms → Measures ↓	**simrank : Algorithm based on nilarity**	**Page ranking using social annotation based on language model**
Description	Rank the paper by comparing the content of query with annotation of page	Rank based on the two strategies i.e. query-annotation similarity and query-document similarity
Input parameters	Papers and query contents	Initial search result list, set of tags and papers
Importance	it uses similarity measure for effective clustering	gives more optimized and accurate result
Limitation	Its efficiency gets affected by the capabilities of the web crawler being utilized	Annotations may contain incomplete and unrelated terms

Table 3. Comparison of algorithms under more than one mining technique

Algorithms → **Measures ↓**	**Future rank : Ranking Articles by Predicting their Future Page Rank**	**Popularity and similarity based page rank algorithm**	**Ranking articles by using citations, authors, journals and time information**
Technique used	Structure mining and content mining	Structure mining and Usage mining	Structure mining and Content mining
description	Ranking is based on the future citation that paper is going to get in future.	The search result list is ranked based on Markov model output and frequency of transition and similarity of papers.	Similar to future rank but uses journal information along with other information
Input parameters	Citations, author, time information	Web sessions (Sequence of pages accessed).	Citations, author, time, information, journals
importance	It combines citations, authors and publication time in an effective way for ranking the article by predicting article's future ranking	It makes prediction method better by analyzing the user's browsing pattern	It takes combination different types of information and time aware weights for ranking the articles, which give more better results
limitation	It does not involve the importance of publication venues such as conferences and journals	This method fails in case of predicting one more step ahead	require compact graph