मयि सर्वमिदं प्रोतं
BSAITM
Since 2001

Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org
Volume 5, Special Issue 2, December (2015)

An Advance Approach of Page Ranking Using Combination of Web Structure Mining and Web Content Mining

Yogita Garg
CSE Dept., B.S.Anangpuria Institute of Technology and Management, Alampur, Faridabad, India

Vinod Jain
Asst. Professor , CSE Dept., B.S.Anangpuria Institute of Technology and Management, Alampur, Faridabad, India

Abstract – The World Wide Web is popular and interactive medium to propagate information today. The web is huge, diverse, dynamic, widely distributed global information service centre. In the highly competitive world and with the broad use of the web in e-commerce, e-learning and e-news, finding user's need and providing useful information are the primary goals of websites owners. Therefore, analyzing user's patterns of behavior becomes increasingly important. Web mining is used to discover the content of the web, the user's behavior in the past, and the web pages that the users want to view in the future. Web mining is used to categorize users and pages by analyzing the user's behavior, the content of pages, and the order of URLs that tend to be accessed in order. Web structure and web content mining play an important role in this approach. Web content mining is extraction and integration of useful data from web page content. Web Structure Mining deals with hyperlink structure of the web. Most of the users rely on search engine to search the web. But the results returned by the search engine are not mostly relevant to user's query and ranking of pages are not efficient according to users requirement. So In this paper few algorithms which uses link structure or web structure mining and few algorithms which uses web content mining have been analyzed for calculating the page rank value of web page. In order to improve the precision of ranking of the web pages, after analyzing the page rank and its various versions, a new algorithm has been proposed in this paper, which uses both web structure mining as well as web content mining as hybrid for calculating the page rank value of WebPages. This gives better and efficient results as compare to others and overcome some limitations of web structure mining based algorithms.

Index Terms – In link, out link, page rank, Search engine, visit of link (VOL), Web Content Mining, Web Structure Mining, Weighted page rank.

1. INTRODUCTION

This World Wide Web (WWW) document plays a starring role for retrieving user requested information from the web resources [9]. It is an enormous, contrary diverse, dynamic and mostly form less data warehouse. As on today WWW is a huge information depository for awareness indication [10]. It has seen an explosive growth. It is estimated that WWW has expanded by about 2000% since its evolution and is doubling in size every six to ten months [1]. In order to retrieve user requested information, search engine plays an important role for crawling web content on different node and organizing them in to result pages so that user can easily select the required information by navigating through the result pages link [11]. This strategy worked well in earlier because, number of resources for user request is limited. Also, it is feasible to identify the relevant information directly by the user from the search engine results. When the internet era increases sharing of resources also increases and this lead to develop an automated technique to rank each web content resource. Different search engine uses different techniques to rank search results for the user query. As the competition and web resource increases, ranking of web content become tedious and dynamic with respect to user query. This also affects user interest on looking for search engines to identify the web content relevant to their needs. Figure. 1 shows the concept of search engines.

This paper is presented at International Conference on Recent Trends in Computer and information Technology Research on 25th& 26th September (2015) conducted by B. S. Anangpuria Institute of Technology & Management, Village-Alampur, Ballabgarh-Sohna Road, Faridabad.

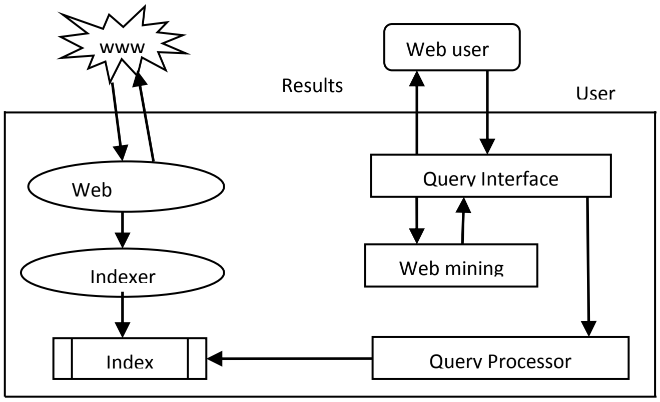

Figure 1 Search Engine Architecture

There are lots of search engines but few like Google, Yahoo, etc. are famous because of their crawling and ranking methodology. So, ranking methodology becomes a very important aspect of web mining in all the three components of search engine (i.e. Crawler, Indexer, Ranking mechanism). Search Engine is used to find the information from [8]. They download, index and store hundreds of millions of web pages. They answer thousands of queries every day. They act like content collector as they keep record of all the information available on WWW. In web search, ranking algorithms play an important role in ranking web pages so that the user could get good result which is more relevant to the user's query [8]. When a user makes a query from search engine, it generally returns a large number of pages in response to user queries. This result-list contains many relevant and irrelevant pages according to user's query. As user impose more number of relevant pages in the search result-list. To assist the users to navigate in the result-list, various ranking methods are applied on the search results. The search engine uses these ranking methods to sort the results to be displayed to the user. In this way user can find the most important and useful result first. As most of the ranking algorithms is either link oriented or content oriented. The page ranking algorithm which use web structure mining doesn't care about user's query, Only link structure of WebPages are considered in calculation of page rank value of WebPages. On the other side the page ranking algorithm which uses web content mining take user's query into account and doesn't care about link structure of WebPages for calculating page rank values of the WebPages. Algorithms which use link structure has mainly many challenges like emphasis on old pages, theme drift, page cheating. Theme Drift- Link structure based algorithms use only link structure means more links to page more important the page is and higher the value of page rank. So the results are independent of the keywords and the user's query this is called problem of theme drift. Cheating of pages- Some site owners insert fake links to WebPages in their website to increase the page rank value of those pages that is called page cheating. In this paper, a page ranking mechanism called "A Combined approach of Page Ranking using Web Structure and Web Content Mining" is being devised for search engines, which works on the basis of Weighted Page Rank based on VOL, which takes in to account the number of visit of links to a web page as well as number of inbound links to a web page and similarity value of a web page with respect to query, which takes in to account the concept of web content mining.

The main purpose of the proposed algorithms is finding more relevant information according to user's query. So, this concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behaviour as well as content of web pages, which reduce the search space to a large scale. The paper is organized as follows: a brief summary of related work is given in section 2; Section 3 describes the proposed algorithm in detail. The result analysis is given in Section 4. Section 5 summarizes the result and draws a general conclusion. Section 6 describes the future scope of this proposed algorithm.

2. RELATED WORK

As the demand of information on the web is increasing day by day, so in order to meet that demand, the search engine has to adopt various techniques to prioritize web pages. It is a great deal of work to rank pages such that it gives user most appropriate results according to its requirement. To make it happen various algorithms have been designed and introduced with different perspective. Some algorithms use link structure of web pages whereas other use content to define relevancy of web pages to user queries. Here are some ranking algorithms discussed with their varying nature of web mining category, working and input parameters.

2.1. Page rank algorithm

Page Rank algorithm is the most commonly used algorithm for ranking the various pages. Page Rank algorithm depends upon link structure of the web pages. The Page Rank algorithm is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. The Page Rank considers the back link in deciding the rank score. If the addition of the all the ranks of the back links is large then the page then it is provided a large rank [1],[6] and page rank score is evenly divided among its outgoing links. A simplified version of Page Rank is given by:

$$PR(u)=c\sum_{v\in B(u)}\frac{PR(v)}{Nv}$$

Where the Page Rank value for a web page u is dependent on the Page Rank values for each web page v out of the set Bu (this set contains all pages linking to web page u), divided by the number Nv of links from page v.

2.2. Weighted page rank

Weighted Page Rank Algorithm is an extension of basic Page Rank Algorithm [2]. It assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among it's out link pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). The popularity from the number of in links and out links is recorded as *Win* (*v*,*u*) and *Wout* (*v*,*u*), respectively. Considering the importance of pages, the original Page Rank formula is modified as

$$PR(u)=(1-d)+d\sum_{v\in B(u)}PR(v)Win(v,u)Wout(v,u)$$

$$Win(v,u)=\frac{Iu}{\sum_{p\in R(v)}Ip}$$

Where Iu and *Ip* represent the number of in links of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*.

$$Wout(v,u)=\frac{Ou}{\sum_{p\in R(v)}Op}$$

Where Ou and *Op* represent the number of out links of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*.

2.3. Weighted page content rank algorithm

Weighted Page Content Rank Algorithm (WPCR) is a proposed page ranking algorithm in which according to a user query a sorted order to the web pages returned by a search engine [4]. WPCR is an algorithm based on the numerical value on which the web pages are given in an order. To calculate the importance of the page, web structure mining is used and how much a page is relevant given by web content mining. The popularity of the page defined by the importance which means how much number of pages is pointing to that particular page. Importance cannot be calculated on the basis of in links only, out links are also to be considered here. The matching of the user query with the particular page shows the relevancy of the page. The page is more relevant if it maximally matched to the user query [3].

$$PR(u)=(1-d)+d[\sum_{v\in B(u)} PR(v)*Win(v,u)*Wout(v,u)*(CW+PW)]$$

- *Probability Weight (PW):* It is the probability of the query terms in the web page. This factor is the ratio of the query terms present in the webpage and the total number of terms in the fired query.

- *Content Weight (CW):* It is the weight of content of the web page with respect to query terms. This is the ratio of the sum of frequencies of highest possible query strings in order and sum of frequencies of all query strings in order. The maximum possible strings are selected in such a way that all such strings represent a different logical combination of words.

2.4. Page rank based on visit of links

Gyanendra Kumar [8]. Proposed a new algorithm in which they considered user's browsing behavior. This concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale so; he proposed an improved Page Rank algorithm. In this algorithm we assign more rank value to the outgoing links which is most visited by users. In this manner a page rank value is calculated based on visits of inbound links. The modified version based on VOL is given in equation:

$$PR(u)=(1-d)+d\sum_{v\in B(u)}\frac{LuPR(v)}{TL(v)}$$

Here d is a damping factor, u represents a web page, B (u) is the set of pages that point to u, PR (u) and PR (v) are rank scores of page u and v respectively, Lu is the number of visits of link which is pointing page u from v. TL (v) denotes total number of visits of all links present on v.

2.5. Enhanced ratio rank: Enhancing impact of in links and out links

Enhanced-Ratio Rank[5] also consider ratio of weight of the in links and weight of out links and visit counts of links by users for calculation of the rank value of particular page. It checks which ratio gives the best result i.e. which ratio of in links weight and out links weight helps to give better relevancy of the web pages. New Enhanced algorithm is given as follows:

$$PR(u) = c \sum_{v\in B(u)} \frac{(Vu*0.7*Win(v,u)+0.3*Wout(v,u))PR(v)}{TL(v)}$$

It uses same parameters as Ratio Rank equation. As in equation 70 percent of the weight of in links and the 30 percent of the weight of the out links is being used because this gives better result as compare to other ratios. By using all three parameters for computing the page rank value of WebPages and taking the best ratio of weight of in links and out links gives the better relevancy of web pages. But the problem of theme drift (some link may not give the search results about the query) still exists in this algorithm.

2.6. Weighted page rank based on visit of links

Simple Sharma and Neelam Tyagi [7] proposed this algorithm. In this algorithm more rank value is assigned to the outgoing links which is most visited by users and received higher popularity from number of in link. Here the popularity of out links is not considered which is considered in the original algorithm. The advanced approach in the new algorithm is to determine the user's usage trends. The user's browsing behavior can be calculated by number of hits (visits) of links. The modified version based on WPR (VOL) is given as:

$$WPRvol(u) = (1-d) + d \sum_{v\in B(u)} \frac{LuWPRvol(v)Win(v,u)}{TL(v)}$$

Here d is a damping factor, u represents a web page, B (u) is the set of pages that point to u, WPRVOL (u) and WPRVOL (v) are rank scores of page u and v respectively, Lu is the number of visits of link which is pointing page u from v. TL (v) denotes total number of visits of all links present on v.

3. PROPOSED MODELLING

In this paper new enhanced page ranking algorithm is presented which exploits hybrid approach for calculating page rank value as it uses both web structure mining as well as web content mining. In this algorithm the importance and relevance of the WebPages is calculated by taking into account the average of weighted rank score of web page, which is calculated using weight of in links, number of visit to the link by users and similarity value of the web pages with respect to the query terms, which is calculated using TF_IDF weight.

The proposed algorithm works on the basis of following steps: Similarity Calculation, Weighted Rank Calculation, and Relevancy Calculation.

3.1. Similarity calculation

Similarity of the document with the query means: what query Terms are present in the document, where they are present and how many times? There are many measures to find out the Similarity between query and the page, and even between two Pages also; but we are using a measure which depends on the Weights of the query terms in the user query and in the document. In information retrieval, documents are ranked with these similarity values. It considers the concept of web content mining. Similarity of a page *p* with reference to the query *q* can be measured with a value called similarity value denoted by *sim (q, p)* generally lying between 0 and 1.

The algorithm for calculating the similarity value is as follows:

Algorithm 1: Calculating the similarity value

Input: User's query, Set of retrieved documents

Output: Similarity Value of retrieved documents

Steps:

```
For each retrieved document
{
  For each word in a page do:-
  {
    1.   Calculate inverse document frequency
         IDF=log $_2$[Total number of pages /
```

```
        number of pages that contain word]

    2.  Calculate term frequency
         TF=Number of times a word appear
         in a page.

    3.  Calculate TF*IDF Matrix that
         contain weight of each word in a
         page ($W_{p,t}$).

    4.  Determine weight of each word in Query
             $W_{q,ti}$=[frequency(ti)/max frequency
                                  of term]* ($W_{p,ti}$)

    5.  Calculate length of page and query as
        Len(p)= Sqrt( $(w_{d,t1})^2+(w_{d,t2})^2+\ldots(w_{d,tn})^2$)
        Len(q)= Sqrt( $(w_{q,t1})^2+(w_{q,t2})^2+\ldots(w_{q,tn})^2$)

    6.  Similarity (p,q)= $\frac{\sum_{i=1}^{n} \mathrm{Wp,ti} * \mathrm{Wq,ti}}{\mathrm{Len(p)*Len(q)}}$
  }

}
.
```

3.2. Weighted page rank calculation

The weighted Rank Calculation considers both forward links and Visit of link to find the rank of the page as WPR (Weighted Page Rank) does. It may be noted that in general, back-links contribute more towards the importance of a page rather than forward-links. It considers the web structure mining. The algorithm for calculating the WPR_{vol} is as follows:

Algorithm 2: Calculating Weighted Page Rank

Input: Web graph, In links, out links, visit of links, set of retrieved pages

Output: Weighted page rank of retrieved documents

Steps:

1. Obtain the web graph from link structure of retreived web pages.
2. Assume initial page rank of all the pages to be 1.
3. Using the equation below calculate the page rank score of all the pages .

$$WPRvol(u) = (1-d) + d \sum_{v \in B(u)} \frac{LuWPRvol(v)Win(v,u)}{TL(v)}$$

Here d is a damping factor, u represents a web page, B (u) is the set of pages that point to u, WPRvol (u) and WPRvol (v) are rank scores of page u and v respectively, Lu is the number of visits of link which is pointing page u from v. TL (v) denotes total number of visits of all links present on v.

3.3. Relevancy calculation

This component is responsible for calculating the relevancy based on similarity value (generated in step 1) and weighted page rank (generated in step 2). This relevancy calculation is purely based on the average of similarity values of the pages with respect to the user query and weighted page rank based on visit of link. However if similarity value of a document is zero, then it will consider the document as not relevant at all and makes it total weight (relevancy) as 0. It may be noted that the pages are ordered according to the new rank as given below:

$$\mathrm{Re}\,levancy(P) = \frac{Sim(q,P) + WPRvol(P)}{2}$$

This type of ranking considers not only the relevance and importance of the page, rather the relevance of its back-links too. Now the user is returned with ranked pages according to relevancy.

4. RESULTS AND DISCUSSIONS

It is almost impossible to get the exact relevant result according to user's query that best match his/her search by using a single algorithm. But the algorithm proposed in this paper take the combination of two algorithm and work as a hybrid to determine exact relevant result.

4.1. Example illustrating the working of proposed algorithm

Content of Page A: new York times
Content of page B: new York post
Content of page C: los angeles times
Query is: new new post

A. For Weighted Page Rank Calculation: The web graph generated for following documents is shown in Figure 2.

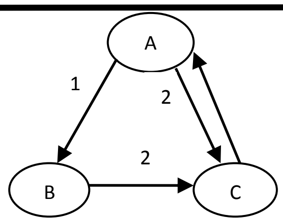

Figure 2.Example illustrating working of WPRvol

Table 1 below shows the iterative calculation for WPRvol

WPRvol(A)	**WPRvol(B)**	**WPRvol(C)**
0.575	0.23145	0.8691
0.51945	0.23145	0.82446
0.50039	0.22088	0.80907
0.49160	0.21996	0.80190
0.575	0.23145	0.8691

Table 1 Iterative calculation of WPRvol

According to the result shown in table 1, the web pages will be displayed in the following order:

Page C > Page A > Page B
Here c gets the higher rank.

B. For similarity calculation: Table 2. Shows the similarity value calculated for each document with respect to query.

Similarity(q,A)	**Similarity(q,B)**	**Similarity(q,C)**
0.34287	0.90755	0

Table 2 Similarity calculation

C. For relevancy calculation: In this step combination of both WPRvol and Similarity value is used.

Table 3 shows the calculated relevancy of each page with respect to query.

Relevancy(q,A)	**Relevancy(q,B)**	**Relevancy(q,C)**
0.4166	0.5635	0

Table 3 Relevancy calculation

According to the result shown in table 3, the web Pages will be displayed in the following order:

Page B > Page A
Here B gets the higher rank

4.2. Precision based comparison of existing system with proposed system

Document retrieval order according to WRPvol : C > A > B

Document retrieval order according to proposed algorithm: B > A

$$Precision = \frac{number\ of\ relevant\ document\ retrieved}{total\ number\ of\ document\ retrieved}$$

Table 4 shows the precision value of both the algorithm:

	WPRvol Algorithm	**Proposed Algorithm**
Precision	2/3=0.667	2/2=1

Table 4 Precision value of both algorithms

Figure. 3 showing the comparison of precision of WPRvol and Proposed Work

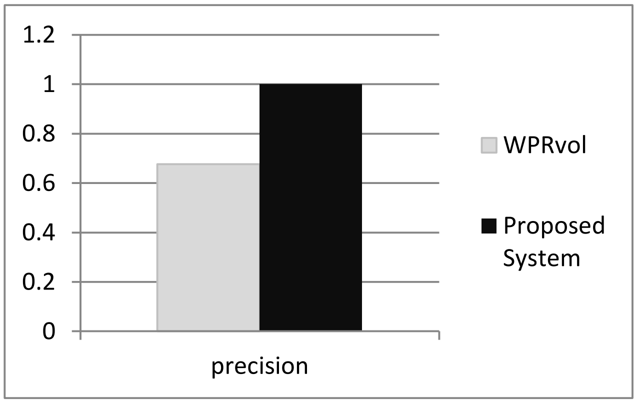

Figure. 3 Precision based comparison of WPRvol and new proposed system

Here it is concluded that the proposed algorithm for relevancy calculation of retrieved web pages gives better precision value (i. e, 1) as compared to previous existing system (i. e, 0.667). All the mathematical equations should be numbered as shown above.

5. CONCLUSION

Web mining is the data mining technique that is used to extract the useful information from web documents. Page rank, weighted page rank, enhanced ratio rank is used in web structure mining. In this dissertation it is focused that these algorithms may not get the required relevant document easily. To solve this problem, we utilize the similarity value to increase the accuracy of web page ranking. This Modified Page Rank Algorithm is based on link structure that consider popularity of in link and number of visit of links as well as web content mining that consider the content of web pages to improve the relevancy of pages.

Finally the proposed algorithm will improve the ranking of a document in case of vast amount of information retrieved by search engine. It also determines how much the retrieved document is useful for the user. Here it is concluded that the proposed algorithm for relevancy calculation of retrieved web pages gives better precision value (i. e, 1) as compared to previous existing system (i. e, 0.667). .

6. FUTURE SCOPE

The future work of this algorithm includes the following: Some improvements in the proposed method can be done by adding some other methods to make the system more robust and the results can be further enhanced by using clustering technique on the above proposed algorithm. It can also improve the relevancy factor to retrieval the Web documents which can further improve the result set.

ACKNOWLEDGEMENT

This work is supported by B.S.Anangpuria Institute of Technology and Management. The author would like to thank Dr. A.K.Sharma Dean PG and Research, Mr. Pawan Bhadana, H.O.D., B.S.Anangpuria Institute of Technology and Management, for their Inspiring discussions. The author would also like to thank the anonymous reviewers for their constructive comments and feedback on this paper.

REFERENCES

[1] Tarun Kumari,Ashlesha Gupta and Ashutosh Dixit,"Comparative Study of Page Rank and weighted Page Rank Algorithm", *International Journal of Innovative Research in Computer and Communication engineering*,ISSN:2320-9801 Vol, 2, Issue 2, February 2014.

[2] W.Xing and A.Gorbani,"Weighted PageRank Algorithm," *Proceedings of the Second Annual Conference on Communication Networks and services Research*, May 2004,pp. 305-314.

[3] Pooja Sharma, Deepak Tyagi and pawan Bhadana,"Weighted Page Content Rank for ordering web Search result",*International Journal of Engineering Science and Technology*, Vol(2).12,2010,7301-7310.

[4] Nidhi Shalya,Shashwat Shukla and Deepak Arora,"An effective Content Based web Page ranking Approach",*International Journal of engineering Science and Technology(IJEST)*,ISSN: 0975-5462 Vol. 4 No. 08 August 2012.

[5] Ranveer Singh and Dilip Kumar Sharma,"Enhanced RatioRank – Enhancing impact Of In Link and Out Link", *Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).*

[6] Page Rank Citation Technique:Bringing order to the Web,January 29,1998.

[7] Neelam Tyagi and simple Sharma,"Weighted Page rank Algorithm based on Number of Visit of Links of web Page",*International Journal of Soft Computing and Engineering(IJCSE)*, ISSN:2231-2307, Volume-2, Issue-3, July 2012.

[8] Sachin Gupta and Pallvi Mahajan ,"Improvement in Weighted Page Rank based on Visits of Links(VOL) algorithm ",*International Journal of Computer and Communication engineering Research (IJCCER)*,volume 2-Issue 3 May 2014 2-Issue.

[9] P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar,"Content Based Ranking for Search Engine",*International MultiConference of Engineering and Computer Scientists 2012*, Vol I, IMECS 2012, March 14-16,2012.

[10] Neelam Tyagi and Simple Sharma,"Comparative Study of various Page Ranking Algorithms in Web Structure Mining", *International Journal of Innovative Technology and Exploring Engineering(IJITEE)*, ISSN:2278-3075,Volume-1,Issue-1,June 2012.

[11] Yogita Garg and Vinod Jain,"A Brief Survey of Various Ranking Algorithms for Web Page Retrieval in Web Structure Mining", *International Journal of Engineering Trend and Technology(IJETT)*,Volume 21 Number 3-March 2015.