ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

Positioning Webpage Using Rank

S. U. Balvir,[1] G. N. Tikhe,[2] L. M. Barapatre[3]

Department of Computer Science & Engineering, DMIETR, Sawangi(M), Wardha

Abstract:- The large information available on the www, web search engine plays a very important role. When user gives any query to search, search engine gives the large amount of pages related to the query. But user wants only useful information. Therefore a ranking mechanism is applied to give the useful result at the top and leaving other result at the bottom. There are varies ranking algorithm. This paper focused on the two ranking algorithms: PageRank and Page Content Rank Algorithm.

Keywords:-WWW; Search engine; Web mining; Web Page Ranking;

1. INTRODUCTION

World Wide Web (WWW) is a vast resource of hyperlinked and heterogeneous information including text, image, audio, video, and metadata. It is estimated that WWW has expanded by about 2000% since its evolution and is doubling in size every six to ten months [1]. With the rapid growth of information sources available on the WWW and growing needs of users, it is becoming difficult to manage the information on the web and satisfy the user needs. Actually, we are drowning in data but starving for knowledge. Therefore, it has become increasingly necessary for users to use some information retrieval techniques to find, extract, filter and order the desired information.

Majority of the users use information retrieval tools like search engines to find information from the WWW. Some commonly used search engines are Google, msn, yahoo search etc. They download, index and store hundreds of millions of web pages. They answer tens of millions of queries every day. They act like content aggregators (Fig. 1) as they keep a record of every information available on the WWW.

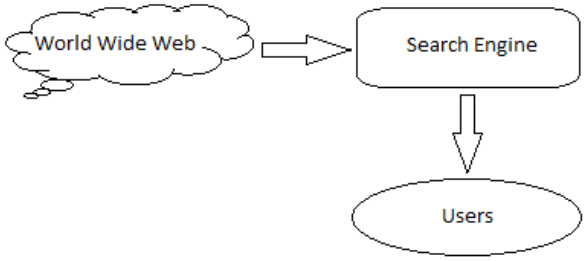

Figure 1. Concept of Search Engine

The most important component of the search engines are shown in Figure2 is a crawler also called a robot or spider that traverses the hypertext structure in the web and downloads the web pages. The downloaded pages are routed to an indexing module that parses the web pages and builds the index based upon the keywords present in the pages. Index is generally maintained alphabetically considering the keywords.

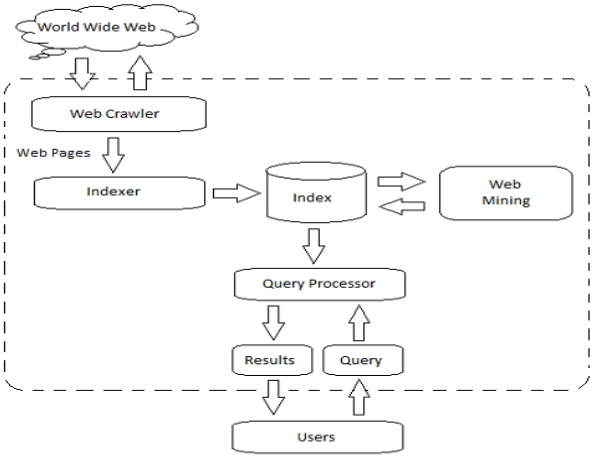

Figure 2. Search Engine Architecture

When a user fires a query in the form of keywords in the search engine, it is retrieved by the query processor component, which after matching the query keywords with the index returns the URLs of the pages to the user.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

But before representing the pages to the user, in back end or in front end is used by most of the search engines to make the user search navigation easier between the search results. Important pages are put on the top and the less important pages in the bottom of the result list. Such kind of mechanism is used by a popular search engine Google that uses the PageRank algorithm to rank its result pages.

This paper describe two ranking algorithms PageRank and Page Content Rank Algorithm. PageRank algorithm. This paper structure as follows: Section 2: contains concept of web mining and their catogories, Section 3: contains description of PageRank Algorithm and Page Content Rank Algorithm, Section 4: finally here concluded with some future suggestion.

2. WEB MINING

Extraction of interesting information or patterns from large databases is called Data Mining. Web Mining [1] is the application of data mining techniques to discover and retrieve useful information (knowledge) from the WWW documents and services. Web mining can be divided into three categories [2,3,4] namely web usage mining, web content mining and web structure mining as shown in figure 3.

2.1 Web Structure Mining (WSM):

Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. This completion takes place through use of spiders scanning the Web sites, retrieving the home page, then, linking the information through reference links to bring forth the specific page containing the desired information. The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps.

2.2 Web Content Mining (WCM):

Web content mining, also known as text mining, is generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. some ranking mechanism (web mining) either This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query. The main uses for this type of data mining are to gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information.

2.3 Web Usage Mining (WUM):

Web usage mining is the third category in web mining Web usage mining is used to discover user navigation patterns and the useful information from the web data present in server logs, which are maintained during the interaction of the users while surfing on the web. Most existing Web analysis tools provide mechanisms for reporting user activity in the servers and various forms of data filtering. Using such tools it is possible to determine the number of accesses to there server and to individual files, the times of visit and the domain names and url's of users.

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the need of Web-based application

3. WEB PAGE RANKING

Since the early stages of the world wide web, search engines have developed different methods to rank web pages. Until today, the occurrence of a search phrase within a document is one major factor within ranking techniques of virtually any search engine. The occurrence of a search phrase can thereby be weighted by the length of a document (ranking by keyword density) or by its accentuation within a document by HTML tags.

Ranking becomes a search engine's most distinguishing process, as this will determine what and how information is displayed to the user. A commonality all search engines share by nature is the organization of pages by relevancy starting first with most relevant and ending with least. The higher a page's rank is, the higher the site's probable relevance will be as perceived by the engine. Every search engine uses its own unique method of determining how pages rank in relation to one another, and these are called algorithms.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

An algorithm is a mathematical formula that will take into consideration dozens of factors that have positive and negative effects on page rank.

Here, a survey of two PageRank algorithms has been done and provide better solution to rank web pages by using advantages from both the algorithms.

3.1. PageRank Algorithm:

PageRank Algorithm is developed by Surgey Brin and Larry Page[5,6] which is used by Google to rank the web pages. This algorithm uses the link structure of the web to determine the importance of web pages.

A simplified version [5,8] of PageRank is defined in Eq. 1:

$$P\text{-}Rank(u) = c \sum_{v \in B(u)} \frac{P\text{-}Rank(v)}{N_v} \qquad \text{Eq.(1)}$$

where:

- u represents a web page.
- B(u) represent a set of web pages that points to u.
- P-Rank(u) and P-Rank(v) represents page rank of u and v respectively.
- Nv denotes the number of outgoing links of page v.
- c is the factor used for normalization.

In this Page Rank algorithm, the page rank score of page (say p) is equally divide among is outgoing links. The values assigned to the outgoing links of page p are used to calculate the ranks of pages pointed to by p. An example given in Fig.4 shows page rank distribution.

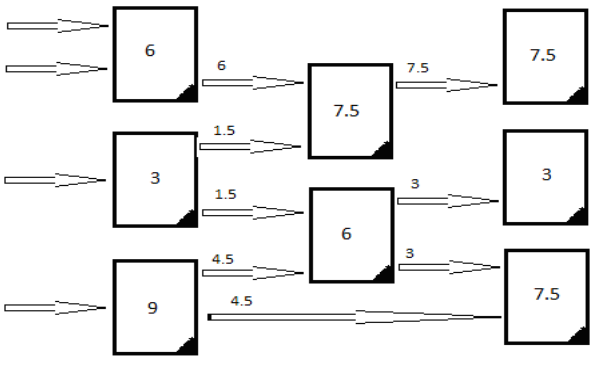

Figure 4. Page Rank Distribution

Later it was observed that not necessary that all user follow the direct link on www. Hence the page Rank was modified as shown in Eq. 2.

$$P\text{-}Rank(u) = (1-d) + d \sum_{v \in B(u)} \frac{P\text{-}Rank(v)}{N_v} \qquad \text{Eq.(2)}$$

where:

- d is the damping factor which is usually set to 0.85 or 0.5.
- d can be thought of as the probability of users following the direct links.
- (1-d) can be thought of as the page rank distribution from non directly linked pages.

Take an example shown in Fig.5 which shows the working of PageRank. This figure shows the hyperlinked structure having three web pages A, B and C.

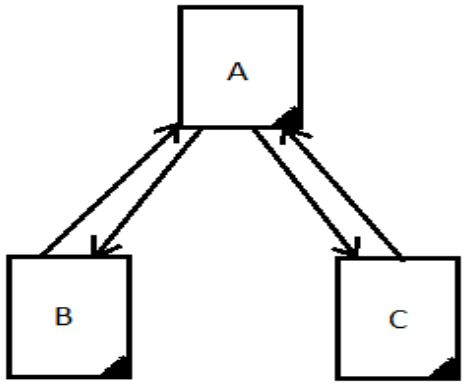

Figure 5. Example

Initially, allocate each page an initial PageRank of 1,although it makes no difference[9] whether we start each page with 1, 0 or 99. Apart from a few millionths of a PageRank point, after much iteration the end result is always the same. Starting with 1 requires fewer iterations for the PageRanks to converge to suitable result thank when starting with 0 or any other number.

By using Eq.2 PageRank for A, B and C are calculated.

P-Rank (A) = (1-d) + d (P-Rank (B)/1+P-Rank (C) / 1)
P-Rank (B) = (1-d) + d (P-Rank (A) / 2)
P-Rank (C) = (1-d) + d (P-Rank (A) / 2)

Consider d=0.85, the page ranks of pages A, B and C becomes:-

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

P-Rank (A) = 1.85,
P-Rank (B) = 0.575,
P-Rank (C) = 0.575.

And after performing 100 iterations, the results are:-

P-Rank (A) = 1.459,
P-Rank (B) = 0.770,
P-Rank (C) = 0.770.

It may be noted that P-Rank(A) > P-Rank(B) = P-Rank(C). Experiments have shown that rank value of a page converges to reasonable tolerance in roughly logarithmic (log n) [5, 6].

3.2. Page Content Rank Algorithm:

The new algorithm proposed for ranking mechanism by Jaroslav Pokorny and Jozef Smizansky[4] which is depends upon the content present in the web pages, called Page Content Rank (PCR). This method performs a number of investigation that seem to be importance for analyzing the content of web pages. Here, importance of a page is depends on the term contained in the web page and the importance of terms is specified by the user giving query q. PCR uses a neural network as its inner classification structure.

3.2.1 Steps Of PCR:-

The PCR method divided into following four steps:-
***(1) Extraction on the basis of term*:-** By using html parse query terms given by user are extracted from each page. An inverted list (index) is built in this step which is used in the step (4).

(2) *Parameters Calculation:-* Statistical parameters such as a term frequency (*TF*) and occurrence positions are calculated. The calculations depend partially on the query *q*, because occurrence positions are calculated relatively to the positions of terms from query *q*.

(3) *Classification on the basis of term*:- Based on parameters from step (2) the importance of each term is determined. As a classifier we use a neural network that is learnt on a training set of terms. Each parameter corresponds to excitation of one neuron in the input level and the importance of a term is given by excitation of the output neuron (there is only one in this neural network) in the time of termination of propagation.

(4) *Finding Page Relevance*:- New page importance are determined in accordance to the importance of terms contained in the pages which is calculated in step (3).

3.2.2 Specification of PCR:-

In PCR, importance of all term present in a page is directly proportional to the importance of that page. In this algorithm, in addition with commonly used aggregation functions like Count, Min, Max, Average, one special function used called Sec_moment shown in Eq. 3.

$$\text{Sec\_moment}(S) = \sum_{i=1}^{n} \frac{X_i^n}{n} \qquad \text{Eq.(3)}$$

where $n = |S|$. Sec_moment is used in this algorithm because it increases the influence of extreme values in the result in contrast to average function.
The following symbols are user in PCR:-

D : The Set of all pages considered for indexing by a search engine,
q : A conjunctive boolean query,
Rq ⊆ D : The set of all relevant pages by the search engine,
Rq,n ⊆ Rq : The set of n top ranked pages from Rq. If n > |Rq|, then Rq,n = Rq.
TF(P, t) : The number of t occurrences in page P(Term Frequency),
DF(t) : The number of pages P which contain the term t (Document Frequency),
Pos(P, t) : The set of positions of t in P,
Term(P, i) : A function returning the term at the i'th position in a function assigning to P.
Term(P, i) = t ≡ i ∈ Pos(P, t).

3.2.3 Parameters deciding the page importance:-

For page importance, it required to calculate the importance of term t, which is denoted by importance(t), and is performed in PCR on the basis of 5+(2*NEIB) parameters, where NEIB denotes the number of neighboring terms included into the calculation. The calculation depends on attribute like set of all pages 'D' indexed by search engine, query 'q' given by user and the number 'n' of pages considered. Further assume a classification function classify() with 5 + (2*NEIB) parameters returning the importance of 't' depending on following parameters.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

***Occurrence frequency of term*:-** This parameter determines the total number of occurrences term *t* in *R*q.

$$freq(t) = \sum_{P \in Rq} TF(P, t) \qquad \text{Eq.(4)}$$

Distance of key terms:- Let QW be the set of all occurrences of terms from Q in all pages in Rq,n, i.e.

$$QW = \bigcup_{t \in Q,\, P \in Rq,\, n} Pos(P, t) \qquad \text{Eq.(5)}$$

Then the distance of t from key terms is the minimum of all distances:

$$dist(t) = \min(\{|it - i| : it \in Pos(P, t) \in \wedge\ i \in QW\})$$

***Incidence of pages*:-** It denoted by Occure(t). this value is a ratio of the number *DF*(*t*) and the total number of pages.

$$Occure(t) = DF(t) / |Rq, n| \qquad \text{Eq.(6)}$$

Frequency in the natural language:- Here assume a database of frequent words and Let *FoL*(*t*) be a mapping from all these words to integers assigning to each word its frequency according to the given database. Then the frequency can be defined as:

$$Common(t) = FoL(t) \qquad \text{Eq.(7)}$$

Term importance:- For the calculation of the rest of parameters we need to know the importance's of all terms from *R*q,n that are determined temporarily as:

importance(t) = classify(freq(t), dist(t), occur(t), common(t), 0, 0,..., 0) Eq.(8)

Synonym class:- This parameter assumes a database having information about classes of synonyms. For each synonym class *S*, calculate an aggregate importance *SC*(*S*) on the base of the importance's of term in the class *S*.

$$SC(S) = Sec\_moment(\{importance(t') : t' \in S\}) \qquad \text{Eq.(9)}$$

This importance is propagated to the term t by another aggregation over all its meanings, i.e.

synclass(t) = sec_moment({SC(St'): t'∈ SENSE(t)}) Eq.(10)

where SENSE(t) contains all meanings t'of t.

Importance of terms neighboring the term t:- The neighboring terms always affect the importance of term i.e. if a term is surrounded by important terms, the term becomes important. It is described by (2*NEIB) parameters, that is an aggregation of the importance of terms surrounding the term t. Let RelPosNeib(t, i), given in Eq.11, be the set of terms which are the ith neighbour of term t in all pages P∈Rq,n, over all occurrences of t. If I < 0 left neighbours are got while i > 0 gives the right ones. The predicate Inside(P, n) is satisfied, if n is an index into the page P. Then,

RelPosNeib(t, i) = ∪P∈Rq,n {Term(P, j + i) : j ∈ Pos(P, t) ∈Inside(P, j + i)} Eq.(11)

and the parameters neib(t, i) for i := −NEIB,−(NEIB−1), … ,−1, 1, … ,NEIB are defined as follows:

neib(t, i) = sec_moment(RelPosNeib(t, i)) Eq.(12)

Based on these parameters the resulted importance of the term t is defined as:

importance(t) = classify(freq(t), dist(t), occur(t), common(t), synclass(t), neib(t,−NEIB),..., neib(t,NEIB)) Eq.(13)

3.2.4 Classification and Importance Calculation of Page:-

In PCR, for classification of pages used a layered neural network, it is denoted by NET, as classification tool. Assume that NET has weights set up from a previous adaptation with a sigmoidal activation function, Assuming the network has 5+(2**NEIB*) neurons in the input layer and one neuron in the output layer. If the calculation of a general neural network *NET* with the input vector ***v*** is denoted as *NET*(***v***) and if *NET*[*i*] is an excitation of the *i*th neuron in the output layer of NET after finishing calculation. Then the *classify*() function can be defined as:

Classify(P1, …, P5+(2*NEIB)) = NET (P1, …, P5+(2*NEIB)) Eq.(14)

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

The importance of a page *P* is calculated as an aggregate value of the importance's of all terms that *P* contains and is written as:

Page importance(P)=sec moment({importance(t): tEP} Eq.(15)

4. CONCLUSION

Web Mining technique is very useful to extract the particular keyword related information from very large amount of data present in WWW. Normally, any search engine gives a large number of result pages in response to users query q. But user always want useful result to be shown on top. Hence, ranking algorithm is applied on the pages which gives the best result to the user. Page ranking algorithms plays a major role to make user search easier.

This paper describe two ranking algorithms PageRank and PCR Algorithm. PageRank algorithm gives ranking to the pages according to the link structure present in the pages as algorithm depends on link structure whereas PCR algorithms gives ranking to the pages according the content present in pages as algorithm depends on content present in pages. In future, the algorithm can be introduced which uses advantages of both the algorithms . The new algorithm will give the better result than both algorithms.

REFERENCES

[1] Raymond kosala, Hedrik Blockeel, "Web Mining Research: A Survey", Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium.

[2] R.Cooley, B.Mobasher and J.Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web". In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997.

[3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", Department of Computer Science and Engineering, University of Minnesota, 200 Union St Se, Minneapolis, MN 55455.

[4] Web Data Mining, http://www.web-datamining.net

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

[6] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Science Department, Stanford University, Stanford, CA 94305.

[7] Jaroslav Pokorny, Jozef Smizansky, "Page Content Rank: An Approach to the Web Content Mining".

[8] A Survey of Google's PageRank: The PageRank Algorithm, http://www.miswebdesign.com/resources/articles/pagerank-2.html

[9] Google's PageRank Explained and how to make the most of it, http://www.webworkshop.net/pagerank.html

[10] http://www.google.com/technology/index.html, Our Search: Google Technology