International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 4 Issue: 1 11-Dec-2014,ISSN_NO: 2321-3337

Analytical survey of Web Page Rank Algorithm

Mrs.M.Usha[1], Dr.N.Nagadeepa[2]

Research Scholar, Bharathiyar University,Coimbatore[1]

Associate Professor, Jairams Arts and Science College, Karur[2]

ABSTRACT— *we use Search Engines to search for information across the Internet. Internet being an ever-expanding ocean of data, their importance grew with every passing day. The diversity of the information itself made it necessary to have a tool to cut down on the time spent in searching. It is very difficult for a user to find the high quality information which he wants to need. When we search any information on the web, the number of URL's has been opened. User wants to show the relevant on the top of the list. So that Page Ranking algorithm is needed which provide the higher ranking to the important pages. In this paper, we discuss the Page Ranking algorithm to provide the higher ranking to important pages.*

Index Terms— Web Mining, Web Usage Mining, Web Structure Mining, Web Control Mining, HITS algorithm, Page Rank.

1, INTRODUCTION

The World Wide Web (Web) is most well-liked and interactive source to broadcast information today. As on today WWW is the largest information repository and set of all nodes which are interconnected by hypertext links. With the quick growth of the Web, users get easily vanished in the rich hyperlink structure. The main aim of website owners is to providing accurate data based on the user's requirement. So, discover the content of the Web pages and retrieving the users' interests from their actions have become gradually more important.

Search Engine Optimization was a term used in the late 90s to show up the importance of a web page's position in results of the search engine. Search engine optimization (SEO) is a well defined process which is used to improve the website rank and also helps to increase traffic to a web site using search engines. SEO process also helps to increase the number of users to a Web site by high ranking in the search results of a search engine. Higher page rank of websites that means that website is more visited by users. The results obtained by a search engines are a combination of large amount of appropriate and inappropriate information. Normally users visit only that website which is top of the lists. SEO is one type technique which helps to find out and get page rank of website from large number of other sites in response to user's search query. So various ranking algorithm such as Page Rank, HITS are available that helps the users to navigate in the results. These ranking method uses by search engine that sort and displayed the result to users. So users can easily find the best result.

The World Wide Web is a very useful and interactive resource of information like hypertext, multimedia etc. When we search any information on the Google, there are many URL's has been opened. The bulk amount of information becomes very difficult for the users to find, extract and filter the relevant information. So that some techniques are used to solve these problems.

Web Mining: Web mining is the application of Data Mining technique to find useful information from web data. With the help of web, we can access multiple data. In the distributed information environment, document or objects are usually linked together to facilitate interactive access to that we can easily access information.

There are some following tasks: [2]

1. **Resource finding:** It is the process which involve to extract data from online or resource available on the web.

2. **Information selection and pre-processing:** The automatic selection and pre-processing of particular information from retrieved web resources and this process transforms the original retrieved data into information.

3. Generalization: It automatically discovers specific patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization.

4. Analysis: It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web.

Web Mining Methodologies

1) Web Content Mining

2) Web Structure Mining

3) Web Usage Mining

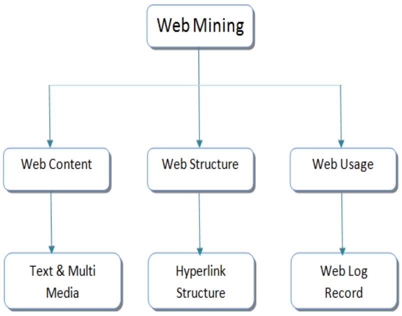

Figure 1 Classified Web Mining

1) **Web Content Mining:** Web Content Mining is the process of retrieving the information from web document into more structure forms. It is related to Data Mining because many Data Mining techniques can be applied in Web Content Mining.

2) **Web Structure Mining:** Web Structure Mining deals with the discovering and modelling the link structure of the web. This can help in discovering similarity between sites or discovering web communities.

3) **Web Usage Mining:** Web Usage Mining deals with understanding user behaviour in interacting with the web site. The aim is to obtain information that may assist web site recognition to better suit the user. The logs include information about the referring pages, user identification, time a user spend at a site and the sequence of pages visited.

There are number of algorithms proposed based on link analysis. But in this paper We are defining: Page Ranking Algorithm.

2, PAGE RANK ALGORITHM

Page Rank Concept : 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.

For the purpose of better search results and especially to make search engines resistant against automatically generated web pages based upon the analysis of content specific ranking criteria (doorway pages), the concept of link popularity was developed. Following this concept, the number of inbound links for a document measures its general importance. Hence, a web page is generally more important, if many other web pages link to it. The concept of link popularity often avoids good rankings for pages which are only created to deceive search engines and which don't have any significance within the web, but numerous webmasters elude it by creating masses of inbound links for doorway pages from just as insignificant other web pages.

Contrary to the concept of link popularity, PageRank is not simply based upon the total number of inbound links. The basic approach of PageRank is that a document is in fact considered the more important the more other documents link to it, but those inbound links do not count equally. First of all, a document ranks high in terms of PageRank, if other high ranking documents link to it. So, within the PageRank concept, the rank of a document is given by the rank of those documents which link to it. Their rank again is given by the rank of documents which link to them. Hence, the PageRank of a document is always determined recursively by the PageRank of other documents. Since - even if marginal and via many links - the rank of any document influences the rank of any other, PageRank is, in the end, based on the linking structure of the whole web.

The Page Rank Algorithm: It was developed at Stanford University by Larry Page and Sergey Brin in 1996. Page Rank is a link analysis algorithm which is used by the Google internet search engine. Page Rank algorithm describes the popularity of web page or website.

This Page Rank algorithm is depend on the link Analysis in which ranking of web page is decided based on outbound links and inbounds links. That means it's totally based on link of WWW and Google uses this algorithm for searching the web pages based on number of hyperlinks such as Inbound and outbound.

Inbound Links: Inbound links are those links that is comes from other site to your website, it is also known as "backlinks". If the backlink comes from an important page, those link will have higher weight than those which are coming from non-important pages. If a link from one page to another page is considered as a vote. Google uses Page Rank to show the important pages move up in the result. Google calculates the importance of the pages from the votes. It can compute the importance of the web page from the votes links.

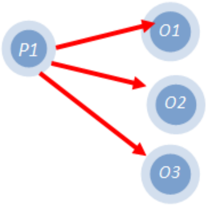

Figure 2 Outbound links pointing to other site

Outbound Links: Outbound links are those links that is pointing to other site from your website and you have more control over these links.

A page has high rank if the other pages with highrank linked to it [7]. It is given by:-

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

- Let A be the page and whose page rank is PR(A).
- Let PR (Ti) is the Pagerank of pages Ti which link to page A,
- C (Ti) is the number of outbound links going out from page Ti and
- d is a damping factor assume to be between 0 and 1 usually 0.85. Sometimes does not click on any links & jumps to another pages at random. It follows the direct links.(1-d) is the probability of jumping off to some random pages; every page has a minimum page rank of (1-d). It follows the non-direct links.

- This damping factor d makes sense because users will only continue clicking on links for a finite amount of time before they get distracted and start exploring something completely unrelated. With the remaining probability (1- d), the user will click on one of the cj links on page pj at random. Damping factor is usually set to 0.85. So it is easy to infer that every page distributes 85% of its original PageRank evenly among all pages to which it points.[8]

To calculate the Page Rank of any Page We required to know the Page Rank of each page that point to it and number of the outbound links from each of those pages.

Let us consider a simple example of three web page A,B and C shown in figure.

1. Page A contains 1 outbound link that is pointing to Page B.
2. Page B contains 2 outbound links that is pointing to Page A and Page C.
3. And Page C contains 1 outbound link that is pointing to Page A
4. The initial page Rank of each page is considered to be 1.

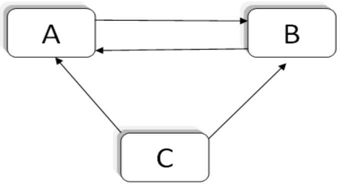

Figure 3Three web pages links between each other

The Page Rank of each page is computed by following equation

PR (A) = 0.2 + 0.4PR (B) + 0.8PR (C)

PR (B) = 0.2 + 0.8PR (A)

PR (C) = 0.2 + 0.4PR (B)

The result of above equation is given

PR (A) = 1.2

PR (B) = 1.0

PR (C) = 0.66

Page rank algorithm is used by famous search engine Google, where the most important web pages are displayed at the top and less relevant pages at the bottom. The Brin and Page applied the citation analysis in web search by treating the incoming links as citations to the web page. Page Rank provides a more advanced way to compute the importance or relevance of web page than simplify the number of pages that are linking to it. For example, if a web page has a link of the Yahoo! home page, it may be just one link but it is a very important one. This page should be ranked higher than many pages with more links but from obscure places. Page rank is an attempt to see how good an approximation of importance can be obtained just from the link structure. The Page rank algorithm provides a more sophisticated method for doing citation counting. The reason that Page rank is interesting is that there are many cases where simple citation counting does not correspond to our common sense notion of importance. [10]

Advantages of Page Rank:[10]

- The important web pages are kept at the top and irrelevant pages are kept at the bottom.
- It is representation of web structure mining which extracts useful information in terms of relevant web pages.
- It calculates the important web pages by incoming and back links and the representation is simple i.e. in graph and linked databases.

Problems of PageRank Algorithm are: [8]

- It is a static algorithm that, because of its cumulative scheme, popular pages tend to stay popular generally.
- Popularity of a site does not guarantee the desired information to the searcher so relevance factor also needs to be included.
- In Internet, available data is huge and the algorithm is not fast enough.
- It should support personalized search that personal specifications should be met by the search result.

3, LITERATURE SURVEY

Analysis of Data Mining Techniques for increasing search speed in web is proposed by "B.Chaitanya Krishna, C.Niveditha, G.Anusha, U.Sindhu, Sk Silar.". In this paper , they are explaining the full detail of PageRank and HITS algorithm and also define the limitations, problems and comparison analysis of both algorithms.[1]

Page Ranking Algorithms for Web Mining is proposed by "Rekha Jain, Dr. G. N. Purohit". In this paper, they are discussing and comparing the PageRank, weighted PageRank and HITS algorithm. They are compared Mining techniques, IP parameters, working, complexity, limitations and search engine of all the algorithms. HITS are used in structure Mining and Web Content Mining. PageRank and Weighted PageRank calculates the score at indexing time and sort them according to importance of page where as HITS calculates the hub and authority score of n highly relevant pages. Complexity of PageRank algorithm is O(log N) where as complexity of Weighted PageRank and HITS algorithms are <O(log N).[2]

Web Mining Research: A Survey is proposed by "Raymond Kosala and Hendrik Blockeel". In this paper, they survey the web mining categories, its methods, applications and some research issues. They also explore the relation between web mining categories and related agent paradigm. [3]

Web Mining: Methodologies, Algorithms and Applications is proposed by "Bussa V.R.R.Nagarjuna, Akula Ratna babu, Miriyala Markandeyulu, A.S.K.Ratnam". They presented an overview of web mining, its methodologies, algorithms and applications. This paper explains methodologies and two most popular algorithms of web mining: HITS and Page Rank. By using web mining algorithms, significant patterns about the user behavior on the web can be extracted and thus improve the relationship between the website and its users. [4]

4, CONCLUSION

This is the survey paper of PageRank algorithm. PageRank is a better approach for calculating the page value which is a numeric value that represents the importance of a page on the web. There is given a formula for calculating the total number of pages which are linked together and counting these links as a vote. Those page will go on the top of the list which has higher number of numeric value or votes.

REFERENCES

[1]Analysis of Data Mining Techniques for increasing search speed in web, "B.Chaitanya Krishna, C.Niveditha, G.Anusha, U.Sindhu, Sk Silar.", International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.1, Jan-Feb 2012 pp-375-383.

[2]Page Ranking Algorithms for Web Mining, "Rekha Jain, Dr. G. N. Purohit", International Journal of Computer Applications (0975 – 8887), Volume 13– No.5, January 2011.

[3] Web Mining Research: A Survey, "Raymond Kosala and Hendrik Blockeel", ACM SIGKDD Explorations, Volume2, Issue-1, 2000.

[4] Web Mining: Methodologies, Algorithms and Applications, "Bussa V.R.R.Nagarjuna, Akula Ratna babu, Miriyala Markandeyulu, A.S.K.Ratnam", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-3, July 2012.

[5] Parveen Rani and Er. Sukhpreet Singh, "An Offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters", International Journal Of Computers & Technology Vol 9, No 1, July 15 ,2013.

[6] Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval, "P Ravi Kumar and Singh Ashutosh kumar", American Journal of applied sciences, 7 (6) 840-845 2010.

[7] D. Achlioptas, A. Fiat, A.R. Karlin and F.McSherry, "Web search via hub synthesis", Proc. Symp. on Foundations of Computer Science, 2001.

[8] N. V. Pardakhe1 and Prof. R. R. Keole "Analysis of Various Web Page Ranking Algorithms in Web Structure Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013.

[9] Laxmi Choudhary and Bhawani Shankar Burdak, "Role of Ranking Algorithms for Information Retrieval".

[10] Deepti Kapila and Prof. Charanjit Singh, "Survey on Page Ranking Algorithms for Digital Libraries", International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 6, June 2014