

# Ranking Based Web Search Algorithms

Akshata D. Deore\*, Prof. R. L. Paikrao\*\*

\* Computer Engg. Department, AVCOE, Sangamner.

\*\* Computer Engg. Department, AVCOE, Sangamner.

**Abstract-** As Internet is rapidly gaining popularity these days, web searching has become more important. A collection of pages where all pages are interlinked looks like a Spider's Web, therefore called as "Web". A search engine makes use of the combination of textual keywords to display the results. When internet user searches some information using search engine, the result contains all the relevant and irrelevant data. As this data is so vast indexing complete web is impractical, so some filtering should be applied to deliver the quality result. A filter mechanism called as "Ranking" is used now days in many popular search engines. This paper gives the idea of Web Search, how it works? And a comparative study of two Ranking based algorithms PageRank and UsersRank for web search.

**Index Terms-** Indexing, PageRank, Search Engine, UsersRank.

## I. INTRODUCTION

Internet has a vast range of information with thousands of web sites and billions of web pages. From this enormous range of information it is difficult to find the information which user wants. Web Search is the process of looking for information stored on the web. When user wants to find some data online, web search helps to get that data within no time. Web searching is the technique where a large set of online data is traversed on user's demand and the results are served to the user. In turn it acts as a filter which filters out some links or web pages depending upon user's request from a vast database of web pages.

Web search can be determined by making the use of Search Engines. A search engine is a computer program which finds the information online when keywords are fed to it by the user and it automates the process of collecting web links [1]. The results produced by search engines are typically prioritized by reference. When the domain comprising requested information is not known, finding resources and information on the web is very difficult without making use of search engine. Search engines regularly crawls the internet looking for information from the web servers and uses indexing to arrange the information, as long as hosts are discoverable to crawlers. Therefore they give latest results and help the people to get the desired information only in one click. Different search engines use different methods and algorithms to crawl or find the information on the internet; the results produced are not same all the time. And regular crawling helps to obtain latest results for the same information.

A portion of web called Invisible Web is used for the hosts who are not discovered by search engines. These invisible webs restrict the crawlers from accessing information they contain. Therefore large part of information from the web remain hidden hence inaccessible to most of the people.

Because of the complex and vast nature of the web the process of finding information and producing results is a bit difficult and hence time consuming. To make this process faster many search engines uses a concept called as "Meta Search". These Meta search engines combine search results produced from different search engines and give a single comprehensive list.

## II. HOW WEB SEARCH ENGINE WORKS?

Web is a collection of pages that are interlinked. This structure is like a Spider Web. Hence a search engine contains its own database where it stores variety of information related to the web. Fig. 1 shows Search engine architecture:

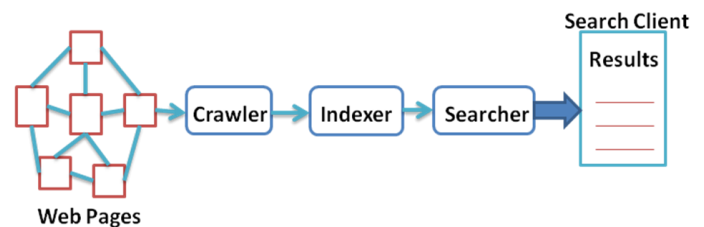


Fig. 1 Search engine architecture

Web search engine is comprised mainly of three modules as given below:

1. Web Crawling
2. Indexing
3. Searching

A program which travels the web from page to page is called as "Web Crawler". To seek out only relevant information from the millions of web pages search engines uses special software Robots known as "Spiders". To establish a list of words obtained from the web sites these spiders roam around the web, this process is known as "Web Crawling" [2]. Many crawling algorithms are available to make successful searches. After gathering the data crawler forwards that data for Indexing.

A complete collection of information given by crawler is then sorted by using Indexers. Indexers are responsible for serving various demands like reading the stored information, uncompressing document and parsing the document [3]. An Anchor file is maintained for storing parsed information of links. Indexing is done to arrange the data in an ordered format like a book index to make it searchable.

Task of a searcher is executing the query requested by the user and producing the results. Searcher looks for the information demanded by user in the index generated by the indexers. The generated result set are then arranged according to user's

preferences and the final list of links is shown in the search engine as an answer to user's query.

### III. RANKING ALGORITHMS

Many ranking algorithms are available for searching the data on the web like PageRank, UsersRank, ObjectRank, etc. In this paper only two algorithms are explained. They are:

#### A. PageRank Algorithm

PageRank was developed at Stanford University by Larry page and Sergey Brin. It is the ranking based algorithms which employees hyperlinks on the web. PageRank algorithm's results are ordered according to pertinence and importance of that page [4]. A rank is calculated for every page on the web by counting total number of votes earned by that page. Using the Markov chain matrix from the vast structure of hyperlinks, static ranking of a web page is formed. Hyperlink from page A to page B interprets one vote from page A to page B [5]. As the PageRank algorithm does not rely on search queries, Rank of a web page is calculated offline and updated after certain duration. It is based on the principle that says when a page contains important incoming links then its outgoing links to other pages are also important [6]. To calculate the Overall page rank previous rank is added to text matching score.

The original PageRank algorithm given by Larry page and Sergey Brin is shown below:

$$PR(A) = (1 - d) + d \frac{PR(T1)}{C(T1)} + \dots + \frac{PR(Tn)}{C(Tn)}$$

Where,

- PR (A) is the Rank of page A,
- PR (Ti) is the PageRank of pages Ti which link to page A,
- C(Ti) is the number of outbound links on page Ti and
- d is a damping factor which can be set between 0 and 1

The PageRank theory holds that even an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor *d*.

PageRank can also be calculated as:

$$PR(X) = \sum_{Y \in B_X} \frac{PR(Y)}{|Y|}$$

Where,

$B_X$  is the numbers of incoming links of page X and  $|Y|$  is the number of outgoing links from Y [7].

#### B. UsersRank Algorithm

When an internet user searches some information online at that time there are chances that the user may bookmark any link or URL if the link is important for him. A bookmark is explicitly generated by the web and it is the place (folder) which stores multiple links when user realises importance of some link. It may be of two types Personal Bookmark or Social Bookmark. A

personal bookmark is stored inside user's system and a social bookmark is shared among users [8]. Bookmark is necessary for some reasons: web searching is dynamic in nature, so even if user finds out the same information after certain amount of time, search engine may produce different results every time. Therefore the link may be lost. Another reason could be it saves the time; instead of searching it again and again user may use it from bookmark's folder [9].

UsersRank algorithm makes use of these bookmarks and produces valuable information for search engines. Here user is treated as a core ingredient for making web search more powerful. It believes in the logic that if user is having some links as bookmarked then those links are actually used by someone hence really valuable and gives effective results for web searches. Main objective of UsersRank Algorithm is to concentrate on the information which is actually referred by number of users thus gives quality search results. Here user is treated as a crawler discovering information using different media and collected information contains a group of URLs visited, gathered and tagged by users.

Every bookmarked entry is considered as a vote given by the user to that page. UsersRank is achieved by summing up total number of votes given by the users to that page.

UsersRank algorithm is shown below:

$$UR(p) = \sum_{q=1}^n R_q(p)$$

Where,

UR(p) is the User Rank of page p. A set R of  $n = |R|$  users is stored in database.  $R_q(p)$  calculates ranking of page p for every  $q^{th}$  user.

#### C. PageRank versus UsersRank

Though PageRank algorithm is popular and widely used, it can be costly because of these reasons: First, web sites are increasing day-by-day hence the size of the web is immensely increasing. It has become a cumbersome task to calculate web matrix, even if rapid functioning computers are used. Second, gradual development of the web allows modifications like adding new pages, deleting old ones, updating links between these pages, etc. So PageRank's quality will be degraded if frequent calculation and modification to a web page's rank is not done. Third, for some web pages more than one rank can be calculated and stored. At that time maintaining multiple ranks for a single page will be a tedious job [10].

UsersRank has many advantages over PageRank. UsersRank supports expansion of web pages comfortably, because of its nature of working. Total number of votes is calculated only using users bookmark's data, so no complexity is involved even if number of web pages grows. As UsersRank generates the results based on the data which is actually used by thousands of users, it gives quality results.

Authers Athanasios Papagelis and Christos Zaroliagis conducted an experiment of 20,000 URLs for comparison. Fig. 2 shows comparison of PageRank versus UsersRank [8]. For each UsersRank range of PageRank and an average PageRank is

shown. According to discovery low UserRank pages frequently have a high PageRank. This means UsersRank results take users point of view into consideration but PageRank may produce results according to predefined methods.

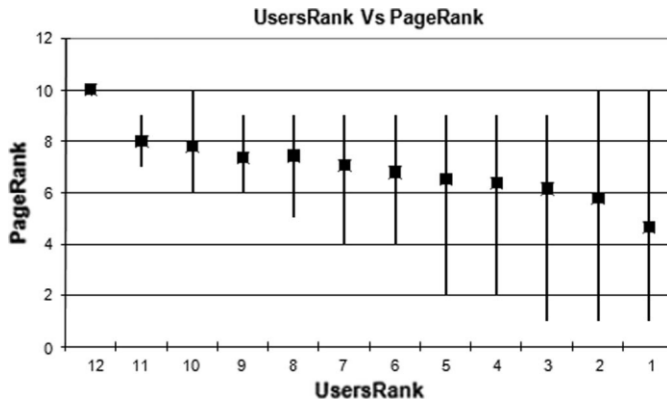


Fig. 2 PageRank versus UsersRank

#### IV. CONCLUSION

In this paper we have studied web searching and how it is done. Also working of two rank based algorithms PageRank and UsersRank is shown.

This work is accompanied by an experimental analysis done by Athanasios Papagelis and Christos Zaroliagis which shows us comparison between PageRank and UsersRank algorithm. Hence we can say that web search is effective when used PageRank algorithm but it can be made more effective by using UsersRank. Also we can produce quality results when we integrate UsersRank with PageRank.

#### REFERENCES

[1] Tauqeer Ahmad Usmani, "A Comparative Study of Google and Bing Search Engines in Context of Precision and Relative Recall Parameter" International Journal on Computer Science and Engineering, ISSN: 0975-3397 Vol. 4 No. 01 January 2012.

[2] Monica Peshave, Kamyar Dezhgosha, "How Search Engines Work and A Web Crawler Application".  
[3] Knut Magne Risvik, "Scaling Internet Search Engines: Methods and Analysis".  
[4] Sara Robinson, "The Ongoing Search for Efficient Web Search Algorithms," From *SIAM News*, Volume 37, Number 9, November 2004.  
[5] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, "Top 10 algorithms in data mining," *Knowl Inf Syst* (2008) 14:1-37 DOI 10.1007/s10115-007-0114-2.  
[6] Neelam Duhan, A. K. Sharma, "A Novel Approach for Organizing Web Search Results using Ranking and Clustering," *International Journal of Computer Applications* (0975 - 8887) Volume 5- No.10, August 2010.  
[7] Catherine Benincasa, Adena Calden, Emily Hanlon, Matthew Kindzerske, Kody Law, Eddery Lam, John Rhoades, Ishani Roy, Michael Satz, Eric Valentine and Nathaniel Whitaker, "Page Rank Algorithm," May 12, 2006.  
[8] Athanasios Papagelis and Christos Zaroliagis, "A Collaborative Decentralized approach to Web Search," *IEEE PART A: Systems and Humans*, Vol. 42, No. 5, September 2012.  
[9] David Abrams, Ron Baecker, Mark Chignell, "Information Archiving with Bookmarks: Personal Web Space Construction and Organization," *CHI 98*, 18-23, April 1998.  
[10] Ali Cevahir, Cevdet Aykanat, Ata Turk, and B. Barla Cambazoglu, "Site Based Partitioning and Repartitioning Techniques for Parallel PageRank Computation," *IEEE Transactions on Parallel and Distributed Systems*, Vol. 22, No. 5, May 2011.

#### AUTHORS

**First Author** – Akshata D. Deore, ME, AVCOE Sangamner, akshatadeore@gmail.com.

**Second Author** – Prof. R. L. Paikrao, ME, AVCOE Sangamner, rahul.paikrao@avcoe.org.