Ranking Techniques in Search Engines

Rajat Chaudhari
M.Tech Scholar
Manav Rachna International University, Faridabad

Charu Pujara
Assistant professor, Dept. of Computer Science
Manav Rachna International University, Faridabad

Abstract - The World Wide Web consists billions of web pages and huge amount of information available within the web pages. To retrieve required information from World Wide Web, search engines perform number of tasks based on their respective architecture. When a user refers a query to the search engine, it generally returns a large number of pages in response to the user's query. To support the ordering of search results according to their importance and relevance to user's query, various ranking algorithms are applied on the search results. this paper gives detailed comparison and analysis of different ranking algorithms: first is the Text Based Ranking, second is PageRank(the Google's algorithm) algorithm; and the last being the Users Rank algorithm.

Keywords: World Wide Web, Information Retrieval, Search Engine, Text Based Ranking ,Page Rank algorithm , Users Rank algorithm.

I. INTRODUCTION

The World Wide Web (www) has already grown up widely and still growing up at a rapid pace. Billions of pages are added every week. The web information on internet is unorganized and unstructured so it is difficult to search without the help of search engines. A search engine[3] start with building a local store for the huge web by downloading the web pages with the help of an agent called *crawler* [10] that visits the URLs looking for the web information to grasp in. this crawler extracts some useful keywords from the entire web. An indexer module builds an index by indexing the information brought by the crawler in the local store. The index consists of keywords and pointers to the location of those keywords on the web.The crawling and indexing form the core functions of a search engine are depicted in Fig. 1. As shown in the figure[2], the searching process begins at the user interface of a search engine where a user fires his/her query by using a language linguistics, may be single or multiple terms, when the user fires the query in to search engine interface than query processor processes the query and match that query to the index created by indexer. After that the ranking module ranks the web pages in order of their relevance or importance by using some ranking algorithm. And the web pages are displayed as the result of users query in decreasing order of their relevance or importance.

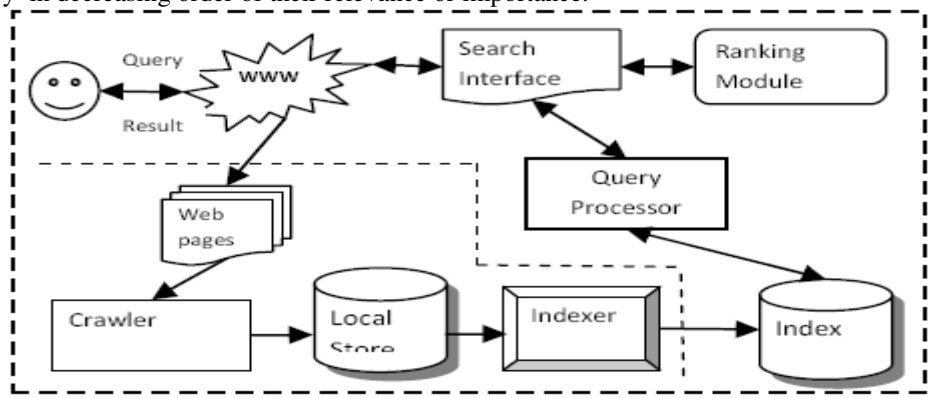


Fig 1: Working Of Search Engine

Search engines uses different Web mining techniques which have their own application areas including search engine optimization, user personalization, usage characterization, ranking of web pages and their rank updation etc.

II. WEB MINING

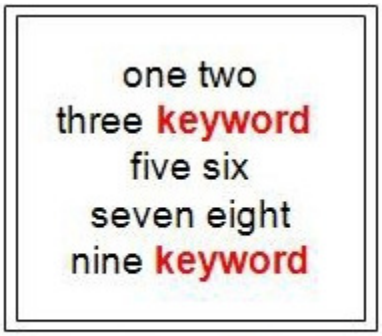
Web Mining [9] is the data mining technique that is used to discover the content of the web, link structure of the web pages and the user's behavior in the past. Web mining can be categorized as Web Structure Mining (WSM), Web Content Mining (WCM) and Web Usage Mining (WUM). WSM is the process of finding out the relationship between web pages by analyzing web structure or web graph. Web graph consists of web pages as nodes and hyperlinks as edges connecting two pages. WCM is responsible for extracting the relevant or useful information from the content of the web pages. WUM identifies user profile and its behavior as recorded in the web log file.

III. RANKING ALGORITHMS

The web page ranking algorithms rank the search results depending upon their relevance to the search query. For this algorithms rank the search results in descending order of relevance to the query string being searched. A web page's ranking for a specific query depends on factors like- its relevance to the words and concepts in the query, its overall link popularity etc. There are three categories of these algorithms viz. text based ranking, PageRank which is based on links and user based ranking.

3.1 Text-Based Ranking

The ranking scheme used in the conventional search engines is purely Text-Based i.e. the pages are ranked based on their textual content, which seems to be logical. In such schemes, the factors that influence the rank of a page are[5]:
a. Number of matched terms with the query string.
b.Location Factors influence the rank of a page depending upon where the search string is located on that page. The search query string could be found in the title of a page or in the leading paragraphs of a page or even near the head of a page [5].
c. Frequency Factors deal with the number of times the search string appears in the page. The more time the string appears, the better is the page ranking[5] .
Most of the times, the affect of these factors is considered collectively. For example, if a search string repeatedly appears near the beginning of a page then that page should have a high rank .
For example we find the keyword density for calculating keyword frequency. Keyword Density is a function, a calculation, of **keyword frequency**. It's calculated as *number of occurrences* divided by *number of words* and is usually expressed as a percentage.



2 / 10 = 0.2 x 100 = 20%

3.2 Page Rank Algorithm

PageRank algorithm is used by Google Search Engine Which is Based On HyperLink Structure Of Web. One of the most important factors that Google uses is PageRank. PageRank is a numeric value that represents how important a page is on the web. Off course PageRank is not the only factor, which decides importance of page, but

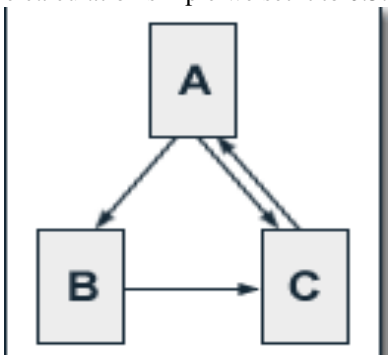
still it is one of them. Sergey Brin and Lawrence Page[1] defined PageRank as: "We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85 …… C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) "

Where:

- PR(A) is the PageRank 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 d is a damping factor which can be set between 0 and 0.85.

EXAMPLE:- We regard a small web consisting of three pages A, B and C, whereby page A links to the pages B and C, page B links to page C and page C links to page A. According to Page and Brin, the damping factor d is usually set to 0.85, but to keep the calculation simple we set it to 0.5.



PR(A) = 0.5 + 0.5 PR(C)

PR(B) = 0.5 + 0.5 (PR(A) / 2)

PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

We get the following PageRank™ values for the single pages:

PR(A) = 14/13 = 1.07692308

PR(B) = 10/13 = 0.76923077

PR(C) = 15/13 = 1.15384615

But As the Size of Web Is Large So We Use The Iterative Method To Compute Page Rank.

Though PageRank algorithm is popular and widely used, it can be costly because of these reasons:

1. Websites are increasing day-by-day hence the size of the web is immensely increasing. It has become a cumbersome task to calculate web matrix.
2. Gradual development of the web allows modifications like adding new pages, deleting old pages, 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.

3.3 UsersRank Algorithm[4] :

UsersRank algorithm makes use of bookmarks and produces valuable information for search engines. When an internet user searches some information online at that time there are chances that the user may bookmark[8] any link or URL if the link is important for him. 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:

$$UR(p)=R1(p)+R2(p)+.....Rn(p)$$

Where,UR(p) is the User Rank of page p. A set R of n = |R| users is stored in database. Rn(p) calculates ranking of page p for every nth user.

Authers Athanasios Papagelis and Christos Zaroliagis conducted an experiment of 20,000 URLs for comparison. Fig. 2 shows comparison of PageRank versus UsersRank[7]. For each UsersRank range of PageRank and an average PageRank is shown.

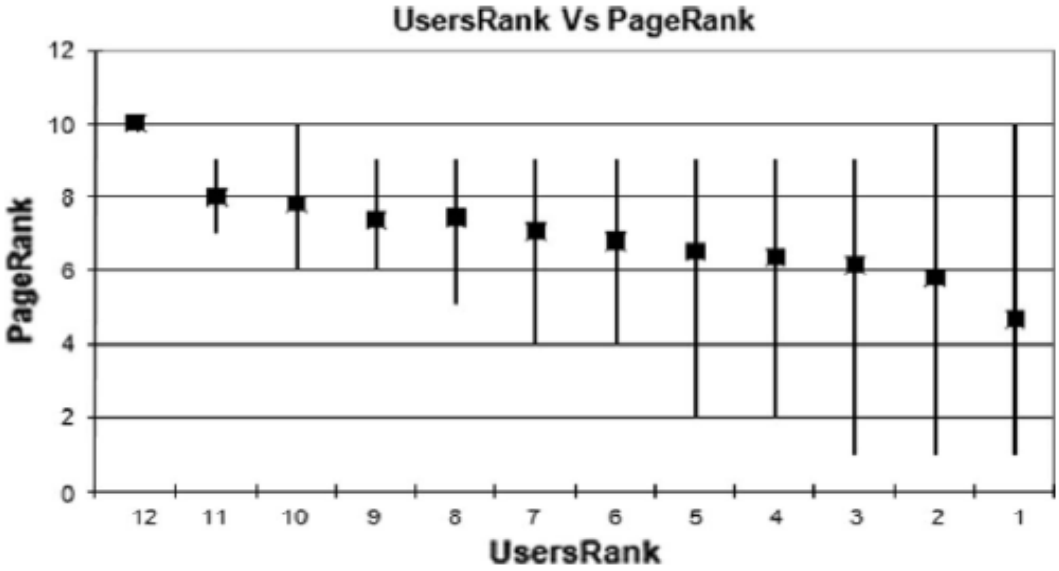


Fig2:PageRank versus usersRank.

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 method.

IV. Comparison between Text Based Ranking ,Page Rank and Users Rank.

parameters	**Text based ranking**	**Page Rank**	**Users Rank**
Description	Based on textual content.	Based on incoming Links	Based on Bookmarks
Input parameters	Number of matched terms with query string or keyword frequency or location of matched query.	Inbound Links of pages	No of bookmarks of page.
Mining Technique used	Web content mining	Web structure mining	Web usage mining
Cost	Lower than page rank and higher than user rank.	highest	lowest
Quality of result	Lowest quality	Higher quality	Highest quality
Relevancy of pages	Completely relevant	no	Partially relevant
Importance of pages	No	yes	yes
Nature of rank	Dynamic (Rank changes with change in content of a page)	Dynamic(Rank changes with link structure)	Dynamic(Rank changes with number of bookmarks of a page)
Rank Distribution	No rank distribution among outgoing links	Ranks are equally distributed to outgoing links	No rank distribution among outgoing links
Advantage	Highest Relevant result	1.Highest important result. 2.pageRank can not be same for 2 or more pages. so it is easy for ranker to decide page ranks.	1.Highest popular result. 2.supports expansion of web pages comfortably, 3.lowest cost
Limitations	1.factors like keyword density can be same for 2 or more pages so it is difficult for rankers to decide which page should be on top 2.it can be easily cheated by keyword stuffing	1.highest cost due to dynamic nature of web 2.can be easily cheated by adding more spamming pages. 3.can be cheated by link farming 4.can be cheated by page stuffing	1.it only checks the bookmarked pages so some other important or relevant pages can be left from entire web.

On the basis of above comparison we have analyzed that text based ranking gives the result only according to the relevancy of the web pages to the user's query .it does not considers the importance of the web pages. On the other hand page rank algorithm gives the result according to the importance of the web pages and does not consider the relevancy factor. But as the Users Rank algorithm gives the result according to numbers of bookmarks of the webpage so it takes user's browsing behavior in to consideration. as the bookmarks are relevant as well as important, we can say that users rank algorithm takes both the factors relevance as well as importance of the web

page in to consideration. Hence users rank algorithm gives better quality results as compared to both text based ranking and Page rank algorithm.
Other advantage of Users Rank over text based ranking and page rank algorithm is its support for dynamic behavior of web i.e. the web pages can be expended comfortably because the votes are calculated according to bookmarks, so no complexity is involved even if number of web pages grows. On the other hand text based ranking and page rank algorithm can be costly because of the two 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 Page Rank's quality will be degraded if frequent calculation and modification to a web page's rank is not done.

V. CONCLUSION

This paper provides detailed study of three ranking algorithms i.e.Text based Ranking,PageRank and UsersRank. UsersRank has many advantages over PageRank and text based Ranking. 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. Hence we can say that web search is effective when used Page Rank algorithm but it can be made more effective by using Users Rank. Also we can produce quality results when we integrate Users Rank with Page Rank.

REFERENCES

[1] Brin S.; Page L. (1998): The Anatomy of a Large-Scale Hyper textual Web Search Engine. Proceedings of 7th International Worldwide Web Conference, pages 107–117.

[2] Tushar Atreja,A.K. Sharma,Neelam Duhan(july2012):"A Comparision Study Of Web Page Ranking Algorithms".International Journal of Advances In Computing And Information Technology.

[3] Google Technology: http://www.google.com/technology/index.html.

[4] Akshata D.Deore,R.L.Paikrao(oct 2012):"Ranking Based Web Search Algorithms" International Journal Of Scientific And Research Publications,volume2 ,Issue10.

[5] World Wide Web searching technique, Vineel Katipally, Leong-Chiang Tee, Yang Yang Computer Science & Engineering Department Arizona State University.

[6] N. Duhan, A. K. Sharma and K. K. Bhatia(2009): "PageRanking Algorithms: A Survey, Proceedings of the IEEE International Conference on Advance Computing.

[7] Athanasios Papagelis and Christos Zaroliagis9 September 2012), "A Collaborative Decentralized approach to Web Search," *IEEE* PART A: Systems and Humans, Vol. 42, No. 5,

[8] David Abrams, Ron Baecker, Mark Chignell(April 1998.), "Information Archiving with Bookmarks: Personal Web Space Construction and Organization," CHI 98. 18-23.

[9] R.Cooley, B.Mobasher and J.Srivastava(1997), "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).

[10] Monica Peshave, Kamyar Dezgosha ,"How Search engine works and A web Crawler Application".

[11] Google and the Page Rank Algorithm, slides by Szekely Endre 2007.01.18.