International Journal of Recent Trends in Engineering & Research
ISSN : 2455-1457 Impact Factor : 3.344

A COMPARATIVE STUDY OF VARIOUS PAGE RANK ALGORITHMS

Sujata V. Jha , Nijal M.Gohil
Gujarat Technological University

Abstract- Nowadays web is growing too fast to match with it. Thus, it has become very important for the sources to give relevant and qualified result. The main aim of this paper is to get the knowledge about various page rank algorithm and to find the optimize result among them. The comparison will be done on the basis of their speed, limitations, benefits and various other parameters.

I. INTRODUCTION

As the content of information is increasing every day. It's a challenge to know about them and be updated everywhere and at every time. Fig 1.shows about the working of a search engine

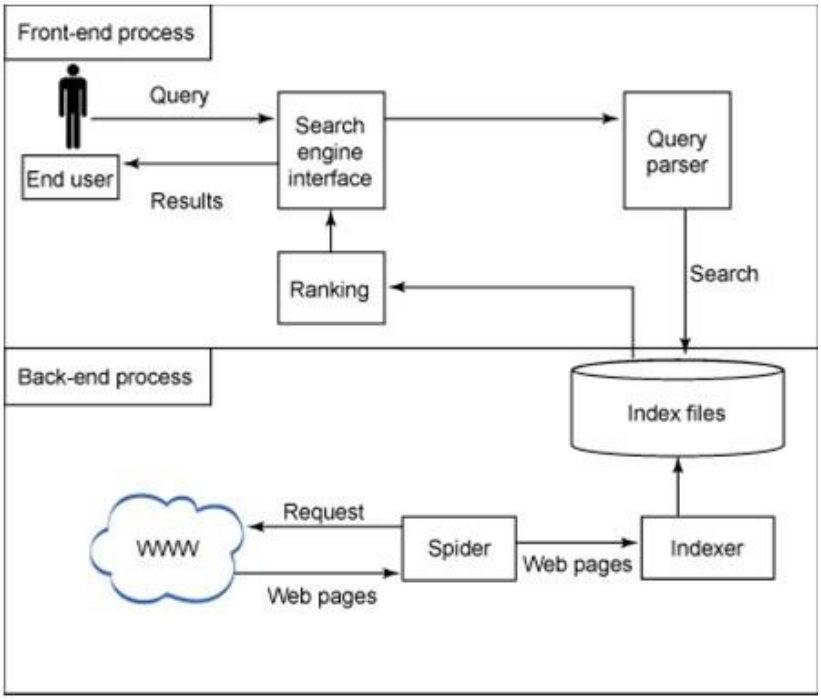

Figure 1-Working of a search engine

In Front end process, a user enters a query, which passes through search engine interface which in turn is parser by a query parser. Query parser searches the query in index files. In Back end process the WWW, from where the web pages are gained or crawled by a spider and it sends the requested web page to the indexer, which will be stored in index files, and through ranking algorithm, the user will get a result.

II. TRADITIONAL ALGORITHMS

PAGE RANK ALGORITHM

Page Rank algorithm [2][3][4] is widely used algorithm for ranking page. Working of the 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. A simplified version of Page Rank is given by:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$

Where the PageRank value for a web page u is dependent on the PageRank 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 L(v) of links from page v. An example of back link is shown in figure 2 below. U is the back link of V & W and V & W are the back links of X.

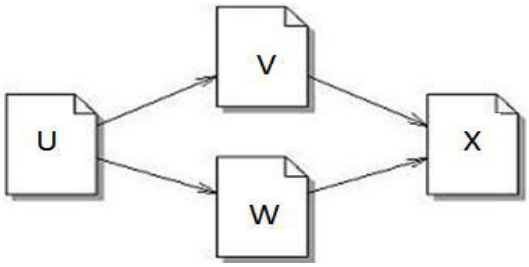

Figure 2 Back link

HITS ALGORITHM

HITS Algorithm is an algorithm that made use of the link structure of the web in order to discover and rank pages relevant for a particular topic.[6] HITS*(hyperlink-induced topic search)* is now part of the Ask search engine (www.Ask.com).

Jon Kleinberg's algorithm called HITS identifies good authorities and hubs for a topic by assigning two numbers to a page: an authority and a hub weight. These weights are defined recursively. A higher authority weight occurs if the page is pointed to by pages with high hub weights.

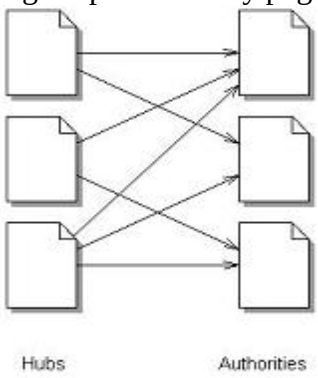

Figure 3. Hits links

HITS is technically, a link based algorithm. In HITS algorithm, ranking of the web page is decided by analyzing their textual contents against a given query.

WEIGHTED PAGE RANK ALGORITHM

Wenpu Xing et. al.[1] discussed a new approach known as weighted page rank algorithm (WPR). This algorithm is an extension of PageRank algorithm. WPR takes into account the importance of both the in links and the out links of the pages and distributes rank scores based on the popularity of the pages. WPR performs better than the conventional PageRank algorithm in terms of returning larger number of relevant pages to a given query. This algorithm provides high value of rank to the more popular pages and does not equally divide the rank of a page among its out-link pages. Every out-link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in links and out links.

PAGE RANKING BASED ON LINKS VISITS (PRLV)

PRLV (Page Ranking based on Link Visits) [5] based on Web Structure Mining and Usage Mining is taking the user visits of pages/links to determine the importance and relevance The weights are calculated for out-links of pages and ranks are computed by taking back-links into account. Calculation of Visits (hits) of links: The hit count of every link is additionally hold on, that is set simply by count the distinct IP addresses visiting the page. Page Ranking based on Visits of Links (VOL) is being devised for search engines, which works on the basic ranking algorithm of Google i.e. PageRank and takes number of visits of inbound links of Web pages into account. 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 reduces the search space to a large scale. The paper also presents a method to find link-visit counts of Web pages and a comparison between VOL with the PageRank algorithm.

EIGENRUMOR ALGORITHM

The EigenRumor algorithm has similarities [7] to PageRank and HITS in that all are based on eigenvector calculation of the adjacency matrix of the links. In the EigenRumor model, however, the adjacency matrix is constructed from agent-to-object links, not page-to-page (or object-to-object) links. Here an agent is used to represent an aspect of human being such as a blogger, and an object is used to represent any object such as a blog entity. Using the EigenRumor algorithm, the hub and authority scores are calculated as attributes of agents (bloggers) and by weighting these scores to the blog entries submitted by the blogger, the attractiveness of a blog entity that does not yet have any in-link submitted by the blogger can be estimated. The implementation experiments of a blog search engine that returns the search result sorted by the scores calculated by this algorithm and evaluated the effectiveness of the ranking by submitting several queries. Links between blog entries are very sparse. Only 1.2% of blog entries have links to the blog entries of others. The aggregation on the agent (blogger) provided the EigenRumor algorithm enables to assign non-zero scores to about 9.3% of blog entries. This greatly improves the usability of blog searches.

DISTANCE RANK ALGORITHM

Distance Rank Algorithm is a novel recursive method based on reinforcement learning which considers distance between pages as punishment, called "Distance Rank" to compute ranks of web pages. [8] The distance is defined as the number of "average clicks" between two pages. The objective is to minimize punishment or distance so that a page with less distance to have a higher rank. The Advantage of this algorithm is that it can find pages with high quality and more quickly with the use of distance based solution. The Limitation of this algorithm is that the crawler should perform a large calculation to calculate the distance vector, if new page is inserted between the two pages.

III. MODERN ALGORITHMS

GOOGLE PANDA [9] is a change to Google's search results ranking algorithm that was first released in February 2011. The change aimed to lower the rank of "low-quality sites" or "thin sites", and return higher-quality sites near the top of the search results. Google Panda is a filter that prevents low quality sites and/or pages from ranking well in the search engine result page.The first Panda update, known as Panda 1.0, was released in February of 2011, affecting nearly 12 percent of all search queries and making it one of the most significant algorithm changes in Google's history. It was intended to weed out low-quality content and reward sites with high-quality, actively updated content.

The next major update was in May of 2014, a rollout known as Panda 4.0, which affected approximately 7.5 percent of all English search queries, adding more features to reward high-quality content and detect instances of low-quality writing.

PANDA 4.1

The rollout began on September 25, 2014, but according to an official Google+ update from Pierre Far, the rollout was a slow, gradual one, planned end in the first week of October. According to Far, the update continues in the tradition of previous Panda updates by identifying low-quality content with more precision. Google has evidently added a few new signals of low quality to its algorithm--but of course, those signals are not public.

PENGUIN

Google Penguin is a code name for a Google algorithm update that was first announced on April 24, 2012.The update is aimed at decreasing search engine rankings of websites [10]. Before, Penguin, Panda came into existence, but because of its faults, it was not up to date. After that other versions of Panda were released. The main target of Google Penguin is spam indexing (including link bombing). Here is a difference between Panda and Penguin. It is true that many site owners find Panda to be much convenient his is named at two lower ranking for low-quality sites. There is only this real difference between Panda and Penguin that helps all website owners to draw and maximize the benefit of achieving higher rankings. Panda searches for low quality websites and it reduces rankings.

Humming bird

The name humming bird comes from precise and fast [11]. The importance of long tailed keywords is basically because of this algorithm. It gives priority to the contextual meaning of the terms used in a query.

The new humming Bird updates works in two ways:-

1. Conversational searches instead of traditional keyword searches.
2. Displays search content right on the search pages which makesit easier for user to pick out the right website.

Hummingbird represents a huge shift in Google's search algorithm from a focus on keyword searches to question-based, conversational, semantic search. Semantic search seeks to provide the best possible results based on what Google knows about you, including your network of contacts, previous searches and social shares, current trends, use of connecting words, and your geographical location. The same search can yield different results to each user, based on their unique circumstances and likely intent. Humming bird uses knowledge graph, a knowledge base used by Google to enhance its search engine search result with semantic search gathered from variety of sources.

IV. COMPARISON OF VARIOUS WEB PAGE RANKING ALGORITHMS

Based on the literature analysis, table 1 comparison is shown on the basis of some parameters such as main technique use, methodology, input parameter, quality of results, importance, limitations

Table 1 Comparison of various web page ranking algorithms

Algorithms	Page rank	HITS	Weighted page rank	PRLV	Eigen rumour	Distance rank
Main technique	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Content Mining	Web Structure Mining

Methodology	This algorithm computes the score for pages at the time of indexing of the pages	It computes the hubs and authority of the relevant pages. It relevant as well as important page as the result.	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided	Calculation of visited links is done .	Eigen rumor use the adjacency matrix, which is constructed from agent to object link not page to page link.	Based on reinforcement learning which consider the logarithmic distance between the pages. .
Input parameter	Back links	Content, Back and Forward links	Back links and Forward Links	Back links	Agent/Object	Forward links.
Quality of results	Medium	Less than PR	More than PR	Medium	Higher than PR and HITS	High
Importance	High. Back links are considered.	Moderate. Hub & authorities scores are utilized.	High. The pages are sorted according to the importance.	User's behavior can b calculated on the basis of pages visited.	High for blog ranking	High. It is based on distance between the pages.
Limitations	Results come at the time of indexing and not at the query time.	Topic drift and efficiency problem	Relevancy is ignored.	These algorithms are missing a crucial factor for how long a webpage is given attention. This can be a great factor to decide the importance of page.	It is most specifically used for blog ranking not for web page ranking as other ranking like page rank, HITS.	If new page inserted between two pages then the crawler should perform a large calculation to calculate the distance vector

Table 2 Comparison of various web page ranking algorithms(contd)

Algorithm	Panda	penguin	Humming Bird
Main technique	Web content mining	Web content mining	Web content mining with deepen semantic search
Methodology	aimed to lower the rank of "low-quality sites" or "thin sites".	aimed at decreasing search engine rankings of websites that violate Google's Webmaster Guidelines.	Conversational search leverages natural language, semantic search, and more to improve the way search queries are parsed.
Input parameter			IP address, geo-data or location
Quality of results	medium	Medium	High
Importance	Moderate. Syndication, User Engagement, Indexation & Keyword Hoarding	Used in webspaming	Conversational search can be done.
Limitation	webspaming	Gaining and loosing traffic	The knowledge graph does not incorporate many languages.

V. CONCLUSION

Depending on the type of algorithms used, it can be decided a definite rank. For a relevant web page. Traditional algorithms have no comparison but the recent ones are also non comparative As, day by day due to technological revolution Algorithm have taken a major role in gathering or releasing the content, information or data After going through comparative study of various algorithms, an efficient algorithm should pass all the critical phases by keeping in mind standard web technology.

REFERENCES

1. Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", in proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.
2. 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.
3. C. Ridings and M. Shishigin, "Pagerank Uncovered", Technical Report, 2002.
4. Sergey Brin and Larry Page, "The anatomy of a Large-scale Hypertextual Web Search Engine", In Proceedings of the Seventh International World Wide Web Conference, 1998.
5. An Empirical Analysis of Page Ranking Algorithms Dr Paras Nath Gupta1, Pawan Singh2, Punit Kr Singh3, Sandeep Chaudhary4, Pankaj P Singh5, Deepak Sinha International Journal of Scientific & Engineering Research, Volume 3, Issue 12, December-2012 ISSN 2229-5518
6. HITS algorithm-Hubs and authorities on the Internet www.math.cornell.edu-winter2009/Roluca Remus
7. Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki,, "The EigenRumor Algorithm for Ranking Blogs", In WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem, 2005.
8. Ali Mohammad Zareh Bidoki and Nasser Yazdani, "DistanceRank: An Iintelligent Ranking Algorithm for Web Pages", Information Processing and Management, 2007.
9. www.huffingtonpost.com/jayson-demers/your-guide-to-googles-panda 4.1
10. en.wikipedia.org/wiki/google-penguin
11. http://www.globalmediainsight.com/blog/hummingbird-google-update hummingbird spawns new search canon.

