A Review of PageRank and HITS Algorithms

International Journal of Advance Research in Engineering, Science & Technology(IJAREST),
ISSN(O):2393-9877, ISSN(P): 2394-2444,
Volume 2,Issue 1, January- 2015

A Review of PageRank and HITS Algorithms

Punit Patel [1]***, Kanu Patel*** [2]

1Assist. Prof, Computer/ I.T Department, Veerayatan Engineering College, Mandvi patelpunitr@gmail.com

2 Assist. Prof, I.T Department, BVM Engineering College, V.V.Nagar kanu.patel@bvmengineering.ac.in

Abstract

The web consists of a huge number of documents that have been published without any quality control. To retrieve necessary information from World Wide Web, search engines carry out number of tasks based on their respective structural design. Various Search engine follow, different algorithm for ranking pages & produce different result. Search engine generally returns a large number of web pages in response to user queries using algorithms. In this paper, we compare two popular web page ranking algorithms namely: HITS algorithm and PageRank algorithm. The paper highlights their variations, respective strengths, weaknesses and carefully analyzes both these algorithms.

Keywords- Page Rank, Ranking Algorithm, Hits Algorithm, Link Analysis

I. 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 has become gradually more important. Higher page rank of websites that means that website is more visited by users. Careful optimization of web sites by Search Engine Optimization that increase the websites visibility in the different search engine Google, Yahoo, Bing and many others. 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. So various ranking algorithm such as PageRank, 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. [9]

In this paper, Page rank Algorithm which works on the number of inbound links and outbound links of web pages. The main goal of the algorithm is to find out relevant information according to users requirement/query. So, this idea is very valuable to exhibit most precious pages on the top of the result list on the basis of user browsing behavior.

II. Data mining over Web

Now a day, the web revolution has had a profound impact on the way we search and find information at home and at work. The web has also become an enormously important tool for communicating idea,

Conducting business and entertainment. Web mining is a data mining technique used to extract information from World Wide Web [2]. Millions of web pages are published every day and millions of are modified or removed. Web pages are written in a different language and provide information in variety of sources such as text, video, audio, image, and animation etc.

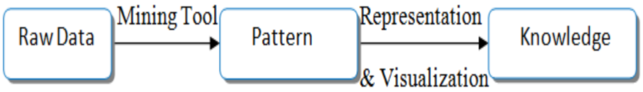


Figure 1 Process of Web Mining

Web Mining Categories

Web mining can be classified into three categories Web Structure Mining, Web Content Mining and Web Usage Mining as depicted in literature [3, 4].

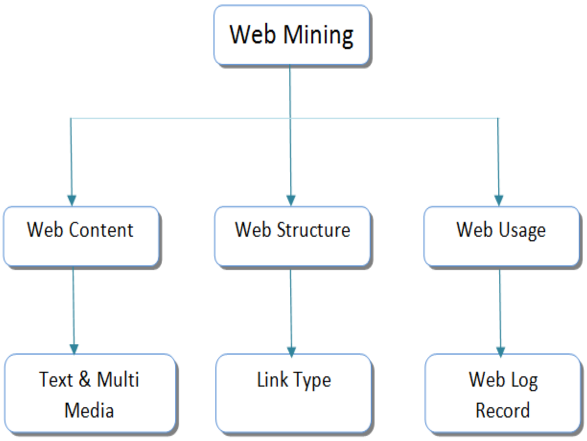


Figure 2 Classified Web Mining

Web Content Mining

WCM is responsible for exploring the proper and relevant information from the contents of web pages. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. [8]

Web Structure Mining

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting two related pages. WSM is used to find out the relation between different web pages by processing the structure of web. Web Structure Mining is useful for extracting structure information from the Web. WSM can be performed at two levels: [8]

1. **Document structure analysis:** deals with the structure of a document such as the Document Object Model.
2. **Link type analysis:** deals with links that may be inter-document or intra-document.

Web Usage Mining

Web Usage mining has been used for various purposes:

1. A knowledge discovery process for mining marketing intelligence information from web data.

2. In order to improve the performance of the website, web usage logs can be used to extract useful web traffic patterns. Web usage mining provides valuable knowledge about user behavior on WWW. One of the major goals of web usage mining is to reveal interesting trends and patterns which can be provide useful information about the user of a system. It includes web server log such as user's IP, referral URL, response status and HTTP request and other.[9]

III. Page Rank

Google has the most well known ranking algorithm called the Page Rank algorithm that has been claimed to supply top ranking pages that are relevant. The Page Rank algorithm was used and enhanced by Lawrence Page and Sergey Brin [5]. 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[6]. 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". Google consider only relevant links point to your site but you cannot control which sites point to your site. If your website content is unique and rich then there are much chances those links will be "dofollow" otherwise links will be consider as "nofollow" [9]

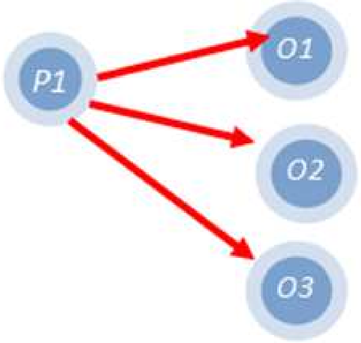


Figure 3. Outbound links pointing to other site

Outbound Links: Outbound links are those links that is pointing to other site from your website and you have more control over these links.[9]

A page has high rank if the other pages with high rank 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.[9]

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.

Example Illustrating functioning of Page Rank

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 contains 1 outbound link that is pointing to Page A

4. The initial page Rank of each page is considered to be 1.

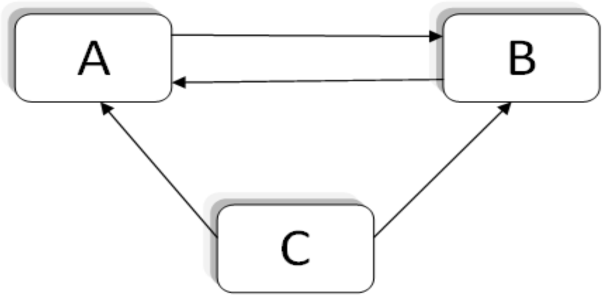


Figure 3. Three web pages links between each other

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

IV. HITS Algorithm

Hypertext Induced Topic Search (HITS) or hubs and authorities is a link analysis algorithm developed by Jon Kleinberg in 1998 to rate Web pages. A precursor to PageRank, HITS is a search query dependent algorithm that ranks the web page by processing its entire in links and out links. Thus, ranking of the web page is decided by analyzing its textual contents against a given query.

When the user issues a search query, HITS first expands the list of relevant pages returned by a search engine and then produces two rankings of the expanded set of pages, authority ranking and hub ranking. In this algorithm a web page is named as authority if the web page is pointed by many hyper links and a web page is named as HUB if the page point to various hyperlinks. The algorithm produces two types of pages: [8]

- ***Authority:*** pages that provide an important, trustworthy information on a given topic
- ***Hub:*** pages that contain links to authorities

Advantages of HITS

1. HITS scores due to its ability to rank pages according to the query string, resulting in relevant authority and hub pages.

2. The ranking may also be combined with other information retrieval based rankings.

3. HITS is sensitive to user query (as compared to PageRank).

4. Important pages are obtained on basis of calculated authority and hubs value.

5. HITS is a general algorithm for calculating authority and hubs in order to rank the retrieved data.

6. HITS induces Web graph by finding set of pages with a search on a given query string.

7. Results demonstrates that HITS calculates authority nodes and hubness correctly. [8]

Drawbacks of HITS algorithm

1. **Query Time cost:** The query time evaluation is expensive. This is a major drawback since HITS is a query dependent algorithm.

2. **Irrelevant authorities:** The rating or scores of authorities and hubs could rise due to flaws done by the web page designer. HITS assumes that when a user creates a web page he links a hyperlink from his page to another authority page, as he honestly believes that the authority page is in some way related to his page (hub).

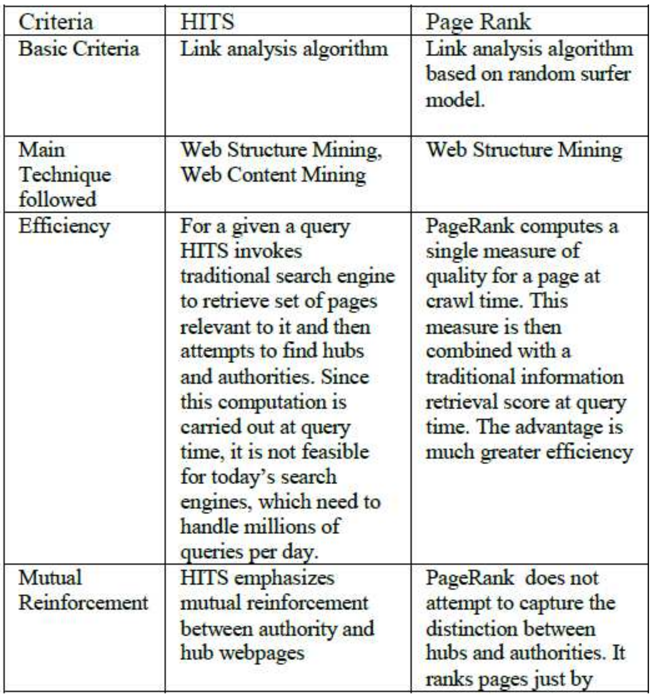
3. **Irrelevant Hubs:** A situation may occur when a page that contains links to a large number of separate topics may receive a high hub rank which is not relevant to the given query. Though this page is not the most relevant source for any information, it still has a very high hub rank if it points to highly ranked authorities.

4. Mutually reinforcing relationships between hosts: HITS emphasizes mutual reinforcement between authority and hub webpages. A good hub is a page that points to many good authorities and a good authority is a page that is pointed to by many good hubs.

5. Topic Drift: Topic drift occurs when there are irrelevant pages in the root set and they are strongly connected. Since the root set itself contains non-relevant pages, this will reflect on to the pages in the base set. Also, the web graph constructed from the pages in the base set, will not have the most relevant nodes and as a result the algorithm will not be able to find the highest ranked authorities and hubs for a given query.

6. Less Feasibility: HITS invokes a traditional search engine to obtain a set of pages relevant to it, expands this set with its inlinks and outlinks, and then attempts to find two types of pages, *hubs* (pages that point to many pages of high quality) and *authorities* (pages of high quality).[8]

V. Comparison of HITS and PageRank



Criteria	HITS	Page Rank
Basic Criteria	Link analysis algorithm	Link analysis algorithm based on random surfer model.
Main Technique followed	Web Structure Mining, Web Content Mining	Web Structure Mining
Efficiency	For a given a query HITS invokes traditional search engine to retrieve set of pages relevant to it and then attempts to find hubs and authorities. Since this computation is carried out at query time, it is not feasible for today's search engines, which need to handle millions of queries per day.	PageRank computes a single measure of quality for a page at crawl time. This measure is then combined with a traditional information retrieval score at query time. The advantage is much greater efficiency
Mutual Reinforcement	HITS emphasizes mutual reinforcement between authority and hub webpages	PageRank does not attempt to capture the distinction between hubs and authorities. It ranks pages just by

VI. CONCLUSION

Now a day Web Mining is a data mining procedure which has become a key part of users. Users generally pay out a lot of time for the search queries and pull out the relevant information from web. We conclude that both page rank and HITS algorithm are different link analysis algorithms that employ different models to calculate web page rank. The Page Ranking algorithms which are an application of web mining play a vital role to easier navigation for users. In this literature review we have discussed about Web Mining and its categorization, beside this we have explained page rank algorithm and how it employ with different concept such as number of users that visit the web pages. And also analyze the page rank of web pages for search engine. However though the HITS algorithm itself has not been very popular, different extensions of the same have been employed in a number of different web sites.

REFERENCES

[1] Parveen Rani, 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*

[2] Tamanna Bhatia," Link Analysis Algorithms For Web Mining ", IJCST Vol. 2, Issue 2, June 2011.

[3] 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.*

[4] Zdravko Markov and Daniel T. Larose, "*Mining the Web: Uncovering Patterns in Web Content, Structure and Usage Data". Copyright 2007 John Wiley & Sons, Inc*

[5] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia.*Page Ranking Algorithms: A Survey* (2009).

[6] Brin, Sergey and Page Lawrence. *The anatomy of a Large-scale hypertextual Web search engine. Computer Networks and ISDN Systems*, April 1998.

[7] Tushar Atreja, A. K. Sharma, Neelam Duhan. *A comparison study of Web Page Ranking Algorithms,IPAJOURNALS*

[8] Nidhi Grover, Ritika Wason, Comparative Analysis Of Pagerank And HITS Algorithms, (IJERT), October – 2012

[9] Punit Patel, Research of Page ranking algorithm on Search engine using Damping factor, Ijaerd, Febuary, 2014