Analysis of HITS Algorithm for Information Retrieval

Khaing Wah Wah Linn
University of Computer Studies (Yangon)
khaingwah2linn@gmail.com

Abstract

As the use of web is increasing more day by day, the World Wide Web is rapidly growing on all aspects. It is the duty of service provider to provide relevant information to the internet user against their query submitted to the search engine. The WWW consists of pages that reference (link to) each other. Interesting pages are referenced by several other pages or interesting pages are referenced by interesting pages. If a page is referenced several interesting pages, it might be itself interesting. This paper present the analysis of one ranking algorithm (HITS algorithm) based on various parameters to find out their advantages and limitations for the ranking of the web pages.
Keyword: HITS algorithm, Information Retrieval, Ranking.

1. Introduction

Today WWW is the huge information repository for knowledge reference. There are a lot of challenges in the Web: Web is large, Web pages are semi structured and so on. Web mining techniques can be used to solve the challenges. Search engines like Google, Yahoo, Web Crawler, Bing etc., are used to find information from the World Wide Web (WWW) by the users. The simple architecture of a search engine is shown in Figure 1. When a user types a query using keywords on the interface of a search engine, the query processor component match the query keywords with the index and returns the URLs of the pages to the user. But before showing the pages to the user, a ranking mechanism is done by the search engines to show the most relevant pages at the top and less relevant ones at the bottom.

2. Related Work

Web mining is the mechanism to classify the web pages and internet users by taking into consideration the contents of the page and behavior of internet user in the past. An application of data mining technique is a web mining, which is used automatically to find and retrieve information from the World Wide Web (WWW). According to analysis targets, web mining is made of three basic branches i.e. web content mining (WCM), web structure mining (WSM) and web usage mining (WUM).

HITS algorithm is used successfully and traditionally in the area of web structure mining [1].

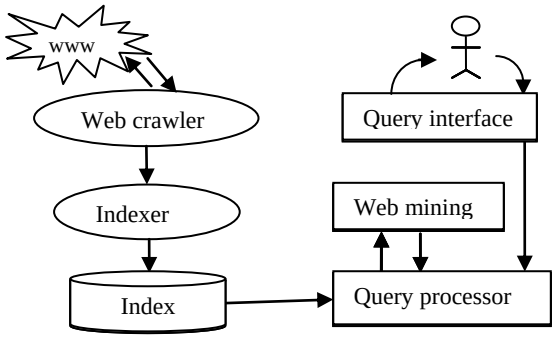

Figure1. Working of Web Search Engine

With the growing number of Web pages and users on the Web, the number of queries submitted to the search engines are also growing rapidly day by day. Therefore, the search engines needs to be more efficient in its processing way and output. Web mining techniques are employed by the search engines to extract relevant documents from the web database documents and provide the necessary and required information to the users.

3. Page Ranking Algorithms

The search engines become very successful and popular if they use efficient ranking mechanisms. Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some ranking algorithms depend only on the link structure of the documents. If the search results are not displayed according to the user interest then the search engine will lose its popularity. So the ranking algorithms become very important [6].

3.1. PageRank algorithm

PageRank algorithm is used by the famous search engine that is Google. This algorithm is the most commonly used algorithm for ranking the various pages. Working of the PageRank algorithm depends upon link structure of the web pages. The PageRank 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 PageRank 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. Therefore, PageRank provides a more advanced way to compute the importance or relevance of a web page than simply counting the number of pages that are linking to it. If a backlink comes from an important page, then that backlink is given a higher weighting than those backlinks comes from non-important pages. In a simple way, link from one page to another page may be considered as a vote [4].

3.2. Weighted Page Rank

PageRank algorithm (WPR) is the modification of the original PageRank algorithm. WPR decides the rank score based on the popularity of the pages by taking into consideration the importance of both the inlinks and outlinks of the pages. This algorithm provides high value of rank to the more popular pages and does not equally divide the rank of a page among it's outlink 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 [4].

3.3. Distance Rank Algorithm

This intelligent ranking algorithm based on reinforcement learning algorithm based on novel recursive method. In this algorithm, the distance between pages is considered as a distance factor to compute rank of web pages in search engine. The main goal of this ranking algorithm is computed on the basis of the shortest logarithmic distance between two pages and ranked according to them so that a page with smaller distance to assigned a higher rank. The Advantage of this algorithm is that, being less sensitive, it can find pages faster with high quality and more quickly with the use of distance based solution as compared to other algorithms. If the some algorithms provide quality output then that has some certain limitations. So the limitation for 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. This Distance Rank algorithm adopts the PageRank properties i.e. the rank of each page is computed as the weighted sum of ranks of all incoming pages to that particular page. Then, a page has a high rank value if it has more incoming links on a page [4].

3.4. EigenRumor Algorithm

The EigenRumor algorithm ranks each blog entry on basis of weighting the hub and authority scores of the bloggers based on eigenvector calculations. So this algorithm enables a higher score to be assigned to a blog entry entered by a good blogger but not linked to by any other blogs based on acceptance of the blogger's prior work. In the recent scenario day by day number of blogging sites is increasing, there is a challenge for internet service provider to provide good blogs to the users. EigenRumor algorithm is proposed for ranking the blogs. The EigenRumor algorithm has similarities to PageRank and HITS in that all are based on eigenvector calculation of the adjacency matrix of the links [4].

3.5. HITS algorithm

HITS algorithm identifies two different forms of Web pages called hubs and authorities. Authorities are pages having important contents. Hubs are pages that act as resource lists, guiding users to authorities. Thus, a good hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many good hub pages on the same subject.

Hubs and Authorities are shown in figure 2. In this a page may be a good hub and a good authority at the same time. This circular relationship leads to the definition of an iterative algorithm called HITS (Hyperlink Induced Topic Selection). HITS algorithm is ranking the web page by using inlinks and outlinks of the web pages. If the web page is pointed by many hyper links, it is called authority and if the page point to various hyperlinks, it is called hub. HITS is a link based algorithm. In HITS algorithm, ranking of the web page is decided by analyzing their textual contents against a given query. After collection of the web pages, the HITS algorithm concentrates on the structure of the web only, neglecting their textual contents [4,5].

In this HITS algorithm, the hub and authority are calculated using the following algorithm.

HITS Algorithm

1. Initialize all weights to 1

2. Repeat until the weights converge:

3. For every hub p∈H

4. Hp = $\sum_{q \in Ip} Aq$

5. For every authority p∈A

6. $Ap = \sum_{q \in Bp} Hq$

7. Normalize

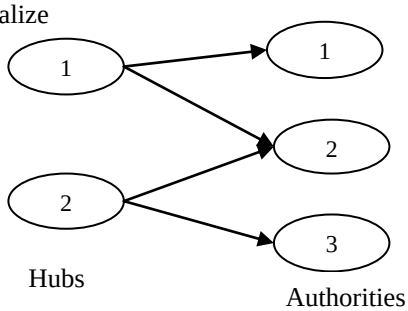

Figure2. Hubs and Authorities

The HITS algorithm treats WWW as a directed graph *G(V,E),* where V is a set of vertices representing

pages and E is a set of edges that correspond to links. There are two main steps in the HITS algorithm. The first step is the sampling step and the second step is the iterative step. In the Sampling step, a set of relevant pages for the given query are collected i.e. a sub-graph S of G is retrieved which is high in authority pages. This algorithm starts with a root set R, a set of S is obtained, keeping in mind that S is relatively small, rich in relevant pages about the query and contains most of the good authorities. The next second step, Iterative step, finds hubs and authorities using the output of the sampling step using equations (3) and (4).

Authority and hub values are defined in terms of one another in a mutual recursion. An authority value is computed as the sum of the scaled hub values that point to that page. A hub value is the sum of the scaled authority values of the pages it points to. Some implementations also consider the relevance of the linked pages.

$$Hp = \sum_{q \in Ip} Aq \qquad (1)$$

$$Ap = \sum_{q \in Ip} Aq \qquad (2)$$

Where H_p represents the hub weight, A_p represents the Authority weight and the set of reference and referrer pages of page p denote with respect *I(p)* and *B(p).* The weight of authority pages is proportional to the summation of the weights of hub pages that links to the authority page. Another one is, hub weight of the page is proportional to the summation of the weights of authority pages that hub link to. Figure 3 shows an example of the calculation of authority and hub scores.

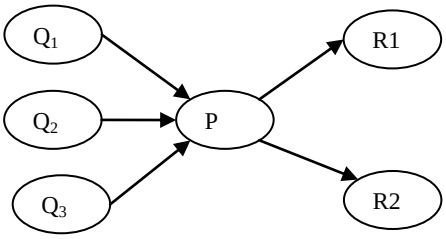

Figure3. Calculation of Hubs andAuthorities

$$AP = HQ1 + HQ2 + HQ3 \qquad (4)$$

Original HITS algorithm has some problems which are given below.

i. High rank value is given to some popular website that is not highly relevant to the given query.
ii. Topic Drift occurs when the hub has multiple topics as equivalent weights are given to all the outlinks of a hub page.
iii. In efficiency: graph construction should be performed on line.
iv. Irrelevant links: Advertisements and Automatically generated links.
v. Mutually effective relationship between hosts: on one site, multiple documents are pointing to document D at another site and retrieve their hub scores and the authority score of D.

4. Conclusion and Further Extension

This paper concludes the introduction of Web mining and the three areas of Web mining. The ranking algorithm provides a definite rank to resultant web pages. Therefore, a typical search engine should use web page ranking techniques based on the specific needs of the users. The main purpose of this paper is to inspect the analysis of HITS algorithm for information retrieval. This is an algorithm of web structure and web content mining technique. It computes the hubs and authority of the relevant pages and it rank pages according to the query topic. But it has two main limitations, topic drift and efficiency. Therefore, HITS's result quality is less than Page Rank algorithm. Existing ranking techniques have limitations particularly in terms of time response, accuracy of results, importance of the results and relevancy of results. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global standards of web technology.

References

[1] Lecture #4: HITS Algorithm - Hubs and Authorities on the Internet file:///C:/Users/user/Desktop/hits/lecture4%28hits%29.htm

[2] "Ranking Hubs and Authorities Using Matrix Functions" Michichele Benzi_, Ernestoestrada, and Christine Klymko.

[3] A. M. Zareh Bidoki and N. Yazdani, "DistanceRank:An intelligent ranking algorithm for web pages" information Processing and Management, Vol 44, No. 2, pp. 877-892, 2008.

[4] Rekha Jain, Dr G.N.Purohit, "Page Ranking Algorithms for Web Mining", *International Journal of Computer application*,Vol 13, Jan 2011.

[5] The PageRank/HITS algorithm, Joni Pajarinen, March 19, 2008.

[6] "Role of Ranking Algorithms for Information Retrieval", Laxmi Choudhary and Bhawani Shankar Burdak, July 2012.

[7] Rekha Jain, Dr G.N.Purohit, "Page Ranking Algorithms for Web Mining", *International Journal of Computer application*,Vol 13, Jan 2011.

[8] "Improving Website Ranking through Search Engine Optimization", Ali H. Al-Badi, Ali O. Al Majeeni, Pam J. Mayhew and Abdullah S. Al-Rashdi.

[9] Modern Improvements and Applications of HITS, Nathaniel Johnson.

[10] "A Comparative Analysis of Web Page Ranking Algorithms", Dilip Kumar Sharma, and A. K. Sharma, India.