ISSN: 0975-766X
CODEN: IJPTFI
Available through Online *Review Article*
www.ijptonline.com

COMPARATIVE ANALYSIS OF PAGERANK AND HITS: A REVIEW

Anudeep Vishwakarma[1], Rishabh Saxena[2], Mohit Awasthi[3] ,Yamuna M[4]*
[1-3]School of Computer Engineering, VIT University, Vellore, Tamil Nadu, India-632014.
[4]School of Advanced Sciences, VIT University, Vellore, Tamil Nadu, India-632014.
Email:anudeep1995@yahoo.co.in

Received on 12-05-2016 *Accepted on 02-06-2016*

Abstract

The World Wide Web is constituted of an ever growing number of web pages. Hence, navigation can become a tall order for the people browsing the Web. Search engines are tools of assistance that yield a number of web pages for a specific query. But identifying relevant web pages from the Web can be a herculean task. The above led to the development of ing algorithms. This paper illustrates how the web pages are linked to among each other. The best method to view the above relationship is as a web graph. Further, the paper discusses the methodologies and working of two famous ing algorithms – Google's and HITS algorithm. It distinguishes the above algorithms from different dimensions highlighting their points of excellence and drawbacks.

Keywords: Comparison, Hits, Pagerank, Web Graphs.

1. Introduction

The World Wide Web (WWW or simply, just Web) has been instrumental in ushering mankind into the Digital Revolution. Majority of people have now made the Web an integral part of their lives, especially in urban geographies. Apart from searching the Web, people also contribute the Web content by publishing articles, videos and images etc. Some people have even made the Web a source of their livelihood by getting paid for publishing. People not only publish but also review content.

The World Wide Web can be viewed as a graph. Such a perspective has been beneficial in understanding the behavior of the Web. People have studied, utilized and exploited the *Web Graph* to prioritize, popularize and publicize certain content over the other. One of the ways of doing so is by using the algorithm. The algorithm is a method to quantify the authority of a document or a page on the Web Graph. Another way of qualifying and ranking a web page is the *Hyperlink Induced*

Topic Search algorithm. It has a different approach to ranking the web pages. These differences are studied to find out which is better than the other in terms of various viewpoints.

Section 2 refreshes various terminologies and definitions from Graph Theory and other related areas that are needed to fully comprehend this paper.

Sections 3 and 4 deal with and HITS respectively.

Section 5 presents a detailed analysis of and HITS algorithms from different viewpoints.

1.1. Terminologies and Definitions

Directed Graphs and its properties[1]

Directed Graph: A directed graph G is a pair G = (V, E), where V is a set of elements called vertices and E is a set of ordered pairs of vertices (u, v) called edges.

In – degree: The indegree of a vertex u is the number of distinct (v, u) $\in$ E in bound u.

Out – degree: The outdegree of a vertex u is the number of distinct (v, u) $\in$ E in out bound u.

Path: A path from vertex u to vertex v is a sequence of edges $(u, u_1), (u_1, u_2), \ldots, (u_k, v)$, where $(u, u_1), (u_i, u_{i+1}, (u_k, v)$ $\in$ E for all i = 1, 2,…, k – 1.

Diameter: A diameter of a graph G = (V, E) is the maximum of all the shortest paths between every possible ordered pair of vertices in V.

Strongly connected component: A strongly connected component (SCC) of a graph G = (V, E) is a set of vertices such that for any pair of vertices u and v in the set there is a path from u to v.

Subgraphs: A graph G' is said to be a subgraph of G if and only if all the vertices and edges of graph G' are also present in graph G.

1.2. Definition of Web Graph

Consider the World Wide Web to be defined as a directed graph. Let the pages or documents available on WWW to be vertices of the directed graph and let the links, which direct one page to another, be edges of the graph. Hence, a Web Graph has been defined. There are other ways of representing web graphs other than a directed graph. One such representation is the S-Node representation by SriramRaghavan and Hector Garcia-Molina[2]. Web Graphs have also been represented by undirected graphs, a perspective whose rigorous treatment has been done by Colin Cooper and Alan Frieze[3].

1.3. Hubs and Authorities

For a query, a web page would be an authority if it contains the relevant information regarding the query. These web pages are referred to by other web pages. A hub is a web page associated with several authoritative web pages. The purpose of a hub is to direct the search engines towards the authorities to give out better and relevant results. They essentially contain links directing to several authoritative web pages.

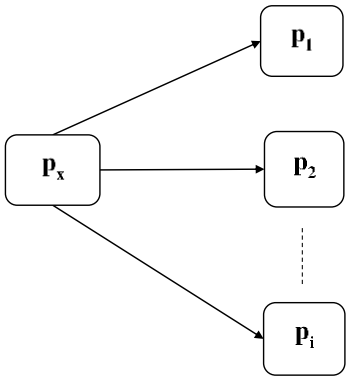

Figure 1: Hub p_x pointing to several authority web pages p_i.

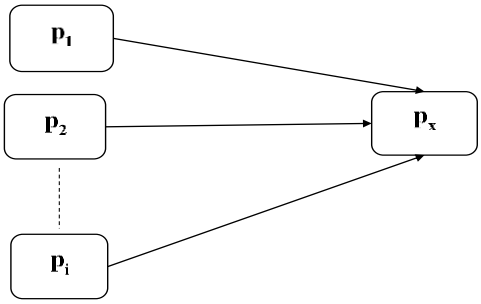

Figure 2: Authority p_x pointed by several hub web pages p_i.

Whenever a query is forwarded to a search engine, two different scores are assigned to it. One of these two scores is called a Hub score and the other score is called the Authority score. To better understand the concept of these two scores, consider the following query made on a search engine – "Best performance laptops available in Indian Market."

There can be some content available on the web which is published by leading laptop brands, which have launched their laptops in India and claim their laptops to be the best in the Indian market. Such pages are called "Authoritative Pages" and have higher Authority Score. There can be other pages on the web which offer information about the laptops based on the experience of the users and customers of particular laptops. These pages will have higher Hub Score. A page having a higher Authority Score is considered to be better if it also has a high hub value. A page having a higher Hub Score is considered to be better if it also has a high Authority Score.

1.4. Link Analysis

There has been an exponential increase in the number of websites and the number of internet users in the last decade. With the world moving at such a fast pace the need to increase the efficiency of the search engines to find the most relevant data within fractions of seconds has been of keen interest. One of the term often used when talking of these algorithms is link analysis. Link analysis, sometimes better known to be mining is the method of using links in order to construct correspondence between entities like deriving relationships between items, their similarity, their differences etc.

1.5. Random Surfer Model

Sergey Brin and Larry Page created the algorithm based on the Random Surfer model[6]. The basis of this model is the user behavior where the user clicks on random links while surfing for the required content[4]. The probability of the user visiting the page is derived from. This probability depends on the number of links present on the page. If the page has more number of links, then the probability of clicking it reduces.

This is the reason why is not passed from one link to another. Instead it is divided by the number of links present on the page. In order to calculate the probability that the random surfer will land on a given page we need to calculate the sum of probabilities of link following to that page. The damping factor d is the probability that the user keeps on clicking links on the pages.

This ranges from 0 to 1. Greater the value of 'd', more are the chances that the surfer will keep on clicking links. Thus (1-d) is the probability that the user stays on the given page and is therefore used as a constant in the algorithm.

1.6. Bow Tie Structure of the Web

According to the classical assumptions, people see net as if the web contains web pages which are completely connected to each other, i.e. it is possible to navigate from any web page on the web to any other web page existing on the web. However, the studies in this field suggest that the possibility to surf between any two random pages is less than one – fourth.

A study was done over 1.5 billion hyperlinks and over 200 million web pages by Researchers from three Californian groups at IBM's Almaden Center, the AltaVista search engine in San Mateo and Compaq Systems Research Center in Palo Alto.

The study suggested that the web actually consists of four different components as shown the diagram[5-6].

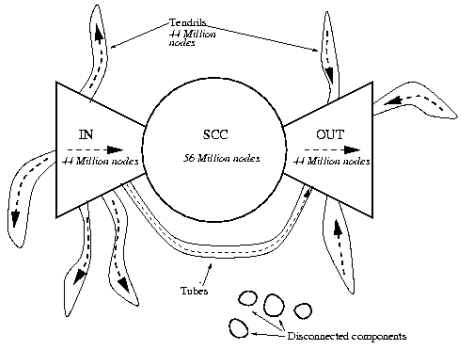

Figure 3: Bow tie structure of web[5-6].

The large circular region, called the core, represents those pages between which a user can surf easily. The segment labeled as 'in' contain the pages which can link (or direct) to the core but we cannot reach them from the pages which are the part of the core segment. The 'out' segment contains those pages which can be reached from the core but cannot redirect the user to the core. These pages may belong to an organization and may have only internal links. Another group of pages exit which can either be connected to 'in' pages or 'out' pages but they are definitely not connected to the core. Apart from these pages, there exist another set of webpages which are not connected to any other group of pages and remain isolated on the web.

2. Algorithm

The basic working of relies on the fact that the most relevant pages for a given keyword are more likely to receive more links or being referred by other web pages. The of a web page can also be thought as the probability of a user ending up or landing up on a particular web page while surfing for a particular phrase or keyword links.

Algorithm is based on link analysis. There has been a lot of criticism about the concept of and its manipulation for falsely influencing the researches are being done in order to identify the false links to pages.

2.1. Approach

The initially divided equally among all the pages in a set or collection. Then starts the process of iteration where the algorithm is found to converge to the theoretical value. This value outputs the probability of a surfer to arrive on a particular page through links. As represents a probability it lies between 0 and 1.

2.2. Algorithm of PageRank

Let us consider 4 web pages: P, Q, R and S. The of all pages is initialized to the same value. The false links i.e. links from page to itself or, or multiple outbound links from one page to another single page, are ignored. Initially considered the sum

of PageRank over all pages to be equal to 1. However, with advancement of PageRank we assume the probability distribution to be between 0 and 1. Thus the initial value of PageRank of each of the page is 0.25 for each A, B, C and D respectively. If links for web page A are present on B, C and D then each link would transfer 0.25 PR to A making its total to be 0.75 in the next iteration.

$$PR(P) = PR(Q) + PR(R) + PR(S) \quad (1)$$

Let us consider page Q to have links to page R and P, and page R having link to page P and page S consisting of links to all the three pages. In the first iteration, page R would transfer all its to the page it links to i.e. P, Page Q would transfer half of its to R and another half to P and Page S would transfer 1/3rd of its existing value to all the three other pages.

$$PR(P) = \frac{PR(Q)}{2} + \frac{PR(R)}{1} + \frac{PR(S)}{3} \quad (2)$$

Thus the PageRank score can be summarized as the sum of PageRank Score divided by the number of links (outbound).

$$PR(P) = \frac{PR(Q)}{L(Q)} + \frac{PR(R)}{L(R)} + \frac{PR(S)}{L(S)} \quad (3)$$

PageRank value for any web page can be expressed as:

$$PR(X) = \sum_{Y \in B_X} \frac{PR(Y)}{L(Y)} \quad (4)$$

i.e. the PageRank value for a page is composed of the sum of PageRank values for each page X contained in the set B_X (the set containing all pages linking to page Y), divided by the number L(X) of links from page X.

3. Hyperlink Induced Topic Search

Jon Kleinberg's Hyperlink-Induced Topic Search, or simply HITS, algorithm was designed and developed to overcome the shortcomings of link analysis based algorithms like PageRank[7]. It came into existence in 1998.

3.1. Problems in Link Analysis based Algorithms:

1. Link analysis do not consider the natural contexts in queries which have a broad sense. For example, the query "which was the most successful IT company in the year 2015" one would expect a company that produces software as people generally consider Information Technology (or IT) companies as companies that produces software. But IT is a broad field which encompasses hardware, software and everything in between. A link analysis based algorithm result in listing of both hardware and software companies, as the algorithm fails to understand the context in which the query was asked.

Similarly, people would expect a listing of fingerprint scanners with the query "cheap biometric devices" but it would result in a list of not only fingerprint scanners but also retina scanners, hand geometry scanners, etc.

2. Some web pages do not include words that are descriptive of the service or information that they provide. For example, one would not find the word "image hosting" on a website like Flickr which hosts images online to share with other people. Similarly, one would not find the word "online social networking" on Twitter which is one of the most popular online social networking site. Since such algorithms depend on the text of the page, they will result in poor listings when queried using the words like the one discussed above
3. Some links in a web page are present only for navigational purposes, like going to the previous or next part of an article. Such links can also be factored in the algorithm resulting flawed listings.
4. Some links in a web page are present for the monetary gains for the owner of the web page such as advertisements which more often than not redirect to irrelevant web pages. Such links can hinder the results of a link analysis based algorithm.

3.2. Approach of HITS

The HITS algorithm categorizes a web page in either of the two categories - Hub or Authority - both which have been discussed in earlier sections. The working of HITS algorithm is highly dependent on the healthiness of both the categories. A healthy authority would be an authority which has more relevant information regarding the subject of the given query than other authorities. A healthy hub would be a hub which has more number of links pointing towards healthy authorities. In other words, a healthy authority has a better authority score than the other web pages. Similarly, a healthy hub has a better hub score than other web pages. Authoritative web pages and their related hubs often overlap. Hence, it is an important part of the algorithm to distinguish between the both. An outcome of such an overlap is the recursive relationship between the hubs and authorities. The relationship is such that:

- A healthy authority is directed to by several healthy hubs.
- A healthy hub directs towards many healthy authorities.

3.3. Algorithm of HITS

The algorithm can be seen as an extension of a text analysis based algorithm. For a certain query, the HITS algorithm acts upon the result of text analysis based algorithm for the query. The result from the text analysis based algorithm is the root

set (R). This root set is essentially a subgraph of the web graph and can also be represented as one. Every web page inside the root set is a may not be an authority of the query. But there must be at least some that may direct to an authority. The algorithm will deal with such pages alone. Hence, to segregate them from the root set, a subgraph of the root set is created that contains links that are directed to or from the root set. This subgraph is the seed or the base set (S). This graph contains the maximum number of authorities for the considered query.

The next step is to define and update the weights of the links from the seed subgraph. The of the above can be mathematically represented as follows:

For every web page there are two associated numeric quantities - weights for hub and authority and let them be notated by $a[p_x]$ and $h[p_x]$ respectively for every page p_x in the seed sub graph containing 'n' pages. These quantities are recursive in nature in a way that one weight is used to define the other due to the inherent co-dependency, as discussed in the earlier section. This recursive nature can be mathematically shown as follows:

$$a[p_i] = \sum_{x=0}^{n} h[p_x], \ \forall p_x \in S \text{ linking towards } p_i \quad (5)$$

$$h[p_i] = \sum_{x=0}^{n} a[p_x], \ \forall p_x \in S \text{ linking from } p_i \quad (6)$$

Since the co-dependent and recursive nature of the above quantities, a balance must be found before moving forward.

Let the hub and authority weight vectors be

$$h = \begin{bmatrix} h[p_0] \\ \vdots \\ h[p_0] \end{bmatrix} \text{and } a = \begin{bmatrix} a[p_0] \\ \vdots \\ a[p_n] \end{bmatrix} \quad (7)$$

For the seed graph S, the adjacency matrix be A and the transpose of the adjacency matrix be A^t, then let the recursive operations be defined as

$$a = A^t \,.\, h \text{ and } h = A \,.u \quad (8)$$

or, upon substituting a and h,

$$a = A^t \,.A \,.u \text{ and } h = A \,.\, A^t \,.\, h \quad (9)$$

Hence we have defined the recursive nature of the weights. This formula is used until a convergence is achieved.

4. Comparison

PageRank and HITS are iterative algorithms that give authorities to web pages on the basis of links directing to them. This being said, both the algorithms have some major differences in the process of granting authority.

Concepts in use: *PageRank uses Random Surfer model. While HITS algorithm is an extension of Link Analysis.*

Field of operation;

PageRank ranks web pages by analyzing the structure of the page. The structure of the page here translates to the number of links and types of links.

HITS algorithm considers both the structure and content of the web page. The content here is the body of the web page which contains paragraphs of text.

Arguments:

PageRank works on the links that direct towards a web page – inward links. While HITS algorithm considers inward links, outward links – links that direct away from a web page; and the content of the web page.

Working:

PageRank algorithm calculates the ranks of pages while indexing them. While HITS algorithm figures the first 'n' relevant web pages.

Confidence:

PageRank has lower confidence levels as it only considers the structure of the web page. HITS have a comparatively higher confidence level while displaying the results owing to the considering of the content of the web page along with its structure.

Area of Application:

PageRank has been an essential part of the Google Search Engine in its formative years. HITS algorithm was developed for IBM's Clever Search Engine.

Association between Hubs and Authorities:

PageRank ignores the hubs and considers only the weights of authorities while ranking a web page. HITS lay importance to both hubs and authorities.

Area of effect:

PageRank studies the entire web as a graph while ranking the web pages. While HITS has a local area of effect as it affects only a subgraph of the web graph.

Query Agnostic:The results of PageRank do not count on the query supplied, as it has already indexed the web pages.

While HITS is a query agnostic algorithm, as it acts upon the initial query results of the initial text analysis

Permanence:

Since both the algorithms are dependent on the structure of the web page, a slight change in the structure results in different results. The change in the web structure can be a manipulation of some links in the web page (be it a hub or an authority).

Application Scenario:

PageRank has been widely used in the field of research for indexing journals. While HITS has a more general area of application.

Susceptibility:

PageRank is less vulnerable to manipulation of links in a localized area such as in a sub graph as it considers the whole web graph before ranking the web pages. While HITS, which ranks pages on the after creating a sub graph of the larger web graph is more vulnerable.

Query Dependency:

The working of PageRank does not depend on the query supplied, as it has already ranked the webpage on the basis on inward links. HITS is highly query dependent, as the sub graph is created from the web graph as a direct result of the query.

Conclusion

This paper analyzed PageRank and HITS algorithms. Though both of these fall under link analysis based algorithms, their approach towards achieving the end results are much different. Links and the relationship between links play a vital role in the functioning of both the algorithms.

PageRank scores over HITS in performance merits such as low vulnerability to local manipulation of links, feasibility and faster results to queries due to indexing. HITS has been known for its accurate computation of hubs and authorities. PageRank has been more popular than HITS due to its inclusion in Google's Search Engine. Whereas there are many deviations and extensions of HITS being deployed by different web sites.

Further work can be undertaken in support of the above analysis pertaining to the two aforesaid algorithms with the help of simulation techniques.

Acknowledgment

We would like to thank Dr. T. Arunkumar, Dean, School of Computer Engineering, for his generous support during the course of writing this paper.

References

1. Narsingh Deo, Graph Theory with Applications to Engineering and Computer Science (Prentice Hall Series in Automatic Computation), Prentice-Hall, Inc., Upper Saddle River, NJ, 1974
2. Raghavan, S. and Garcia-Molina, H., 2003, March. Representing web graphs. In Data Engineering, 2003. Proceedings. 19th International Conference on (pp. 405-416). IEEE.
3. Cooper, C. and Frieze, A., 2003. A general model of web graphs. Random Structures & Algorithms, 22(3), pp.311-335.
4. Blum, A., Chan, T.H. and Rwebangira, M.R., 2006, January. A random-surfer web-graph model. In Proceedings of the Meeting on Analytic Algorithmics and Combinatorics (pp. 238-246). Society for Industrial and Applied Mathematics.
5. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J., 2000. Graph structure in the web. Computer networks, 33(1), pp.309-320.
6. Arasu, A., Novak, J., Tomkins, A. and Tomlin, J., 2002, May. PageRank computation and the structure of the web: Experiments and algorithms. In Proceedings of the Eleventh International World Wide Web Conference, Poster Track (pp. 107-117).
7. Page, L., Brin, S., Motwani, R. and Winograd, T., 1999. The PageRank citation ranking: bringing order to the web.
8. Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.S., 1999. The web as a graph: measurements, models, and methods. InComputing and combinatorics (pp. 1-17). Springer Berlin Heidelberg.

Corresponding Author

Yamuna M*,

Email: *myamuna@vit.ac.in*