Page Ranking Algorithms – A Comparison

K. R. Srinath

Associate Professor, Department of Computer Science, Pragati Mahavidyalaya Degree and PG college, Hanuman Tekdi, Koti, Hyderabad, Telangana, India

---

Abstract - *The growth in the number of websites has been increasing tremendously over the years and the data over the web has been increasing accordingly. Retrieving the required information from the web thereby fulfilling the needs of the web user has become a challenging job for website owners. This paper looks into the insights of the various ranking algorithms and their comparative study.*

Key Words: HITS, PageRank, Weighted PageRank, Web Structure, Web mining, Web content, Web usage.

1. INTRODUCTION

The web is huge, diverse, and dynamic. Retrieving of the required web page on the web, efficiently and effectively, is becoming a challenge. Whenever a user wants to search the relevant pages, he/she prefers those relevant pages to be at hand. The bulk amount of information becomes very difficult for the users to find, extract, filter or evaluate the relevant information. This issue raises the necessity of some technique that can solve these challenges. Most of the web information is in semi-structured form [1] and extracting knowledge from such semi-structured data is difficult.

Today's search engines are plagued by various problems:

- The abundance problem(99% of info of no interest to 99% of people)
- Limited coverage of the web(internet sources hidden behind search interfaces)
- Largest crawlers cover less than 18% of all web pages.
- Limited query interface based on keyword oriented search.
- Limited customization to individual users.

There are 3 vital components in a search engine: Crawler, Indexer and Ranking mechanism.

The Crawler is also called as a robot or spider that navigates the web and downloads the web pages. The downloaded pages are being transferred to an indexing module that parses the web pages and erect the index based on the keywords in individual pages. An alphabetical index is normally sustaining using the keywords. When a query is being floated by a user, it means the query transferred in terms of keywords on the interface of a search engine, the query mainframe section examine the query keywords with the index and precedes the URLs of the pages to the client. But before presenting the pages to the client, a ranking mechanism is completed by the search engines to present the most relevant pages at the top and less significant ones at the substructure. It makes the search outcomes routing easier for the user. In this regard web mining and ranking mechanism becomes very significant for effective information retrieval.

Web mining:

Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from web data.

Web mining subtasks:

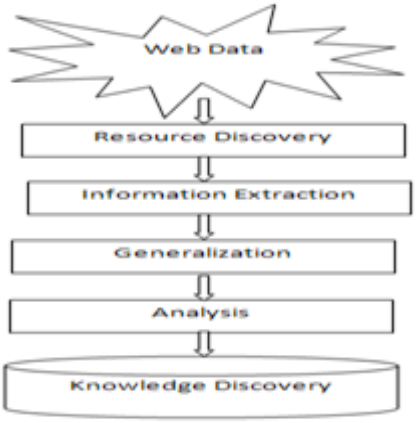

Fig.1 Web mining subtasks

The subtasks of web mining is consists of the following phases as shown in Fig. 1

Resource finding deals with retrieving the intended documents. *Information selection or Preprocessing* which selects and preprocesses the specific information from selected documents. *Generalization* which discovers general patterns within and across web sites and *Analysis* which performs validation and interpretation of mined patterns.

Web mining types:

Web mining is divided into the following 3 types as shown in Fig. 2

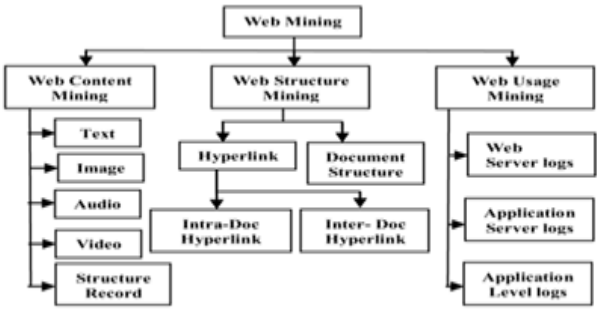

Fig. 2 Web mining Types

Web Content mining:

Web content mining is the process of extracting useful information from the contents of web documents. It includes extraction of structured data from web pages, identify, match and integration of semantically similar data, opinion extraction from online sources, and concept hierarchy, ontology, or knowledge integration. Web content mining is the analogue of data mining techniques for relational databases since we can expect to find similar types of knowledge from unstructured data residing in web documents. The content data consist of text, images, audio, video or structured records.

Web Usage mining :

Web usage mining analyses the transaction data, which is logged when users interact with the web. Web usage mining is sometimes referred to as log mining, because it involves mining the web server logs. Web server logs, which is maintaining an account of each user browsing activity. Web servers automatically, generate large data stored in server referred as logs containing information about the user profile, access pattern for pages, and so on. The world's largest portal like Yahoo, MSN, and so on, needs a lot of insights from the behavior of their user's web visits. Web usage mining collects the data from web log records to discover users' access patterns of web pages. This can provide information that can be used for efficient and effective web site management and user behavior.

Web Structure mining:

Web structure mining focuses on analysis of the link structure of the web and one of its purposes is to identify more preferable documents. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting between two related pages. Web page can also be organized in a tree-structures format, based on the various HTML and XML tags within the page. Technically, web content mining mainly focuses on the structure of the inner document, while web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, web structure mining will categorize the web pages and generate the information, such as the similarity and relationship between different web sites. The goal of web structure mining is to generate structural summary about the web site and web page.

LINK ANALYSIS ALGORITHMS

Web mining technique provides the additional information through hyperlinks where different documents are connected. We can view the web as a directed labeled graph whose nodes are the documents or pages and edges are the hyperlinks between them. This directed graph structure is known as web graph.

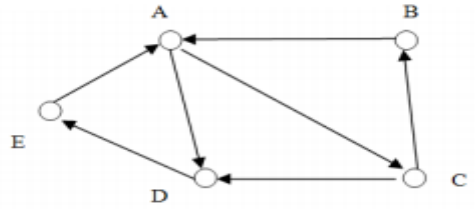

Fig.3 An example of a web graph

There are several algorithms proposed based on link analysis. Three important algorithms PageRank [2], Weighted PageRank [3] and HITS (Hyper-link Induced Topic Search) [4] are discussed below. There are number of algorithms proposed based on link analysis.

A. PageRank Algorithm

PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best-known. PageRank was named after Larry Page, [5] one of the founders of Google. PageRank is a way of measuring the importance of website pages. It is considered the basis for all modern Search Engines. The underlying assumption is that more important websites are likely to receive more links from other websites.

According to Google PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. It Ranks pages based on the number of backlinks pointing to them. The algorithm assigns pages a Total PageRank based on the PageRanks of the Backlinks pointing to the page. The links to a page can be categorized into the following types: Inbound links which are links into the given site from outside so from other pages. Outbound links which are links from the given page to pages in the same site or other sites and Dangling links which are links that point to any page with no outgoing links.

The PageRank of a web page is calculated as a sum of the PageRanks of all pages linking to it(its incoming links),

divided by the number of out links on each of those pages(its outgoing links).

$$PR(A)=(1-d)+d\sum_{i=1}^{n}\frac{PR(Ti)}{C(Ti)} \quad [6]$$

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 1. It depends on the number of clicks, usually set to 0.85
n is the number of inlinks of page A.

Following is a simplified example of the PR algorithm. Consider web graph shown in Fig. 4

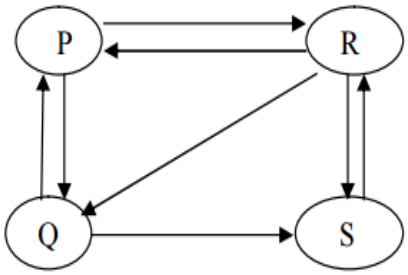

Fig. 4 Example of a Web graph with inlinks and outlinks

Page, P being referenced by pages Q and R. Q, R has 2,3 outlinks. Then PageRank value of the page P is given as:

$$PR(P)=1-d + d(PR(Q)/2 + PR(R)/3)$$

The PageRank algorithm does not rank the whole website, but it is determined for each page individually. Furthermore, the PageRank of page A is recursively defined by the PageRank of those pages which link to page A.

Iterative Method of Page Rank

In iterative calculation, each page is assigned a starting page rank value of 1. These rank values are iteratively substituted in page rank equations to find the final values. In general, much iteration could be followed to normalize the page ranks.

The PageRank algorithm can be iteratively applied as:

1) Initially let Page rank of all web pages is one.
2) Calculate page ranks of all pages by using above formula.
3) Repeat step 2 until values of two consecutive iterations match.

Advantages:

- Since it pre computes the rank score it takes less time and hence it is fast.
- It is more feasible as it computes rank score at indexing time not at query time.
- It returns important pages as Rank is calculated on the basis of the popularity of a page.

Disadvantages:

- The main disadvantage is that it favors older pages, because a new page, even a very good one, will not have many links unless it is part of an existing web site.
- Relevancy of the resultant pages to the user query is very less as it does not consider the content of web page.
- Other problems exists in the form of Dangling links which occurs when a page contains a link such that the hypertext points to a page with no outgoing links.
- It leads to Rank sinks problem occurs when in a network pages get in infinite link cycles.
- Dead Ends are possible ie., pages with no outgoing links.
- Another problem in PageRank is Spider Traps. A group of pages is a spider trap if there are no links from within the group to outside the group.
- If you have circle references in your website, then it will reduce your front page's PageRank.

B. Weighted Page Rank

The weighted page rank algorithm3 was proposed by Wenpu Xing and Ghorbani Ali. Weighted PageRank algorithm is an extension of the PageRank algorithm. Weighted Page Rank allocates a higher rank values to more significant pages instead of dividing the rank value of a page evenly among its outgoing linked web pages. Each outgoing link gets a value proportional to its significance.

A brief explanation of weighted page rank algorithm is given below. In the weighted page rank algorithm, more important (popular) web pages are assigned larger page rank values. The popularity of a web page depends on the number of its inlinks and outlinks and each web page gets a proportional page rank value. The popularity of each page can be obtained using the in and out weights, as given below:

$$W^{in}_{(u,v)} = \frac{I_u}{\sum_{p\in r(v)} I_p} \quad [7]$$

$$W^{out}_{(u,v)} = \frac{O_u}{\sum_{p\in r(v)} O_p} \quad [8]$$

Here r (v) is the set of all Web pages that have inlinks from node v (reference page list of page v). These weights depends on the number of inlinks and outlinks of page u and the sum of the number of inlinks and outlinks of all reference pages of page v, respectively. The initial page rank for each of the n Web pages is given by PR_0 = (PR_0 (1), PR_0 (2),... PR_0

(n)) and their value is set as 1. The formula for computing the weighted page rank of Web page v is given by

$$PR(v)=(1-d)+d\sum_{v\in B(u)} PR(v)\cdot W^{in}_{(u,v)}\cdot W^{out}_{(u,v)} \quad [9]$$

Where B (u) is the set of all web pages that point to u and d denotes the damping factor.

Advantages:

- Quality of the pages returned by this algorithm is high as compared to PageRank algorithm.
- It is more efficient than PageRank because rank value of a page is divided among it's outlink pages according to importance of that page.

Disadvantages:

- As this algorithm considers only link structure not the content of the page, it returns less relevant pages to the user query.

C. HITS Algorithm

Kleinberg developed a WSM based algorithm named Hyperlink-Induced Topic Search (HITS) which presumes that for every query given by the user, there is a set of authority pages that are relevant and accepted focusing on the query and a set of hub pages that contain useful links to relevant pages/sites including links to many authorities. Thus, fine hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many fine hub pages on the same subject. Hubs and Authorities are shown in Fig. 5.

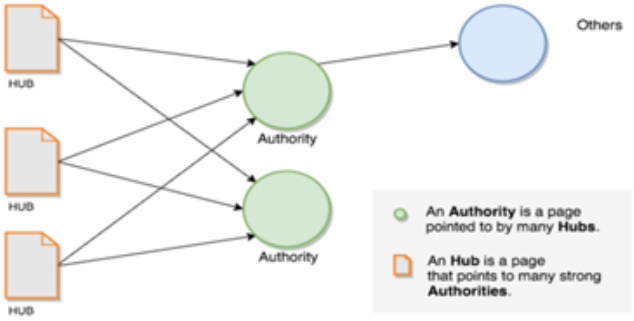

Fig. 5 Hubs and Authorities

Kleinberg states that a page may be a good hub and a good authority at the same time. This spherical relationship leads to the definition of an iterative algorithm called HITS (Hyperlink Induced Topic Search). 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 match up to links.

There are two major 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 influence pages. This algorithm starts with a root set R, a set of S is obtained, keeping in mind that S is comparatively small, rich in relevant pages about the query and contains most of the good authorities. The second step, Iterative step, finds hubs and authorities using the output of the sampling step using:

$$H_p=\sum_{q\in I(p)} A_q \quad [10]$$

$$A_p=\sum_{q\in B(p)} H_q \quad [11]$$

Where H_p is the hub weight, A_p is the Authority weight, I(p) and B(p) denotes the set of reference and referrer pages of page p. The page's authority weight is proportional to the sum of the hub weights of pages that it links to it; similarly, a page's hub weight is proportional to the sum of the influence weights of pages that it links to. Fig. 6 shows an example of the calculation of authority and hub scores.

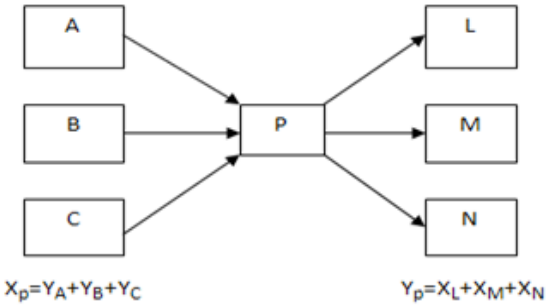

Fig. 6 Calculation of hubs and Authorities B

Advantages of HITS

- HITS scores due to its ability to rank pages according to the query string, resulting in relevant authority and hub pages.
- The ranking may also be combined with other information retrieval based rankings.
- HITS is sensitive to user query (as compared to PageRank).
- Important pages are obtained on basis of calculated authority and hubs value.
- HITS is a general algorithm for calculating authority and hubs in order to rank the retrieved data.
- HITS induces Web graph by finding set of pages with a search on a given query string.
- Results demonstrate that HITS calculates authority nodes and hubness correctly.

Drawbacks of HITS algorithm

- Since HITS is a query dependent algorithm the query time evaluation is expensive.
- 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).
- 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.
- HITS emphasize mutual reinforcement between authority and hub web pages. 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.
- 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.
- 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).

CONCLUSIONS

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. Page Rank and Weighted Page Rank algorithms are used in Web Structure Mining to rank the relevant pages. The standard search engines usually result in a large number of pages in response to users' queries, while the user always desires to get the best in a petite time. The page ranking algorithms, which are an application of web mining, play a major character in making the user search navigation easier in the results of search engine. The PageRank and Weighted Page Rank algorithm give importance to links rather than the content of the pages, the HITS algorithm anxieties on the content of web pages as well as links. Page Rank and Weighted Page Rank algorithms are used in Web Structure Mining to rank the relevant pages. After going through exhaustive analysis of algorithms for ranking of web pages against the various parameters such as methodology, input parameters, relevancy of results and importance of the outcome, it is concluded that on hand techniques have limitations particularly in terms of time response, accuracy of results, importance of the outcome and relevancy of results. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global principles of web technology.

REFERENCES

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

[2] S. Brin, and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[3] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[4] J. Kleinberg, "Authoritative Sources in a Hyper-Linked Environment", Journal of the ACM 46(5), pp. 604-632,199

[5] S. Brin, and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

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

[7] Jaganathan B, Kalyani D. Category-based page rank algorithm. International Journal of Pure and Applied Mathematics. 2015 Aug; 101(5):811-820.

[8] Jaganathan B, Kalyani D. Penalty-based page rank algorithm. ARPN Journal of Engineering and Applied Sciences. 2015 Mar; 10(5):2000–3.

[9] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM. 1999 Sep; 46(5):604–32.

[10] S. Brin, and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[11] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.