CS561-Map Reduce and Big Data
Google's Pagerank and web search

Sanskar Gupta
B18140

***Abstract*—Web mining is a technique for categorizing individuals and sites by examining user behavior, page content, and the order in which URLs are often viewed. HITS and PageRank are two popular page ranking algorithms. To increase the performance of these approaches, several algorithms have been devised. The WPR is a variation of the traditional PageRank algorithm.**

1. Introduction

Web content mining, web structure mining, and web use mining are the three types of web mining. Web mining is a technique for discovering web content, historical user activity, and web pages that people desire to visit in the future. WSM categorizes web pages and produces related patterns based on the topology of hyperlinks. WCM is concerned with extracting meaningful information from the web material. WUM determines user-profiles and the behaviour of users as captured in the weblog file.

2. Background

With the fast expansion of the Internet, it is becoming increasingly challenging to provide relevant, high-quality websites. The reasons for this are because certain online sites are not self-descriptive, and some links are only for navigation. As a result, locating relevant pages through a search engine that depends on online data is extremely challenging.

In WebStructure Mining, PageRank is a widely utilised algorithm. It analyses the connections to determine the relevance of the sites. Google created PageRank, which is named after Larry Page, the company's co-founder and president.

Google utilises Page Metrics to rank web pages based on keywords and other criteria like title tags and keywords.

3. The HITS Algorithm

Hypertext Induced Topic Selection (HITS) assigns a score to a webpage based on the number of inbound and outbound links. Authorities are websites that have a lot of links pointing to them, whereas hubs are websites that point to a lot of links.

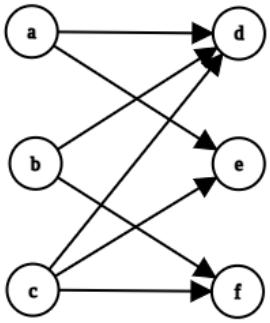

Figure 1. Hubs - {a, b, c} & Authorities - {d, e, f}

Scores are calculated in a mutually reinforcing manner: a strong authority is one that is linked to multiple high-scoring hubs, and a popular hub is one that links to several popular authorities.

Let's define some of the terms for page p,

1) Authority score, a_p
2) Hub score, h_p
3) Referrer pages, B(p)
4) Reference pages, I(p)

The following formula is used to determine hub and authority scores:

$$a_p = \sum_{q \in B(p)} h_q$$

$$h_p = \sum_{q \in I(p)} a_q$$

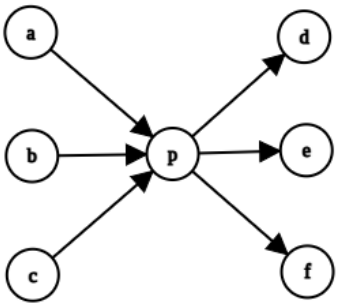

Figure 2. Authority score $a_p = h_a + h_b + h_c$ & Hub score $h_p = a_d + a_e + a_f$

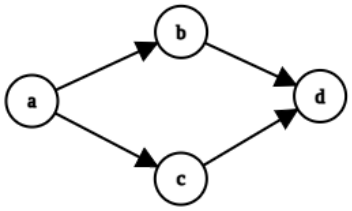

Figure 3. Page a is a backlink to pages b and c, and pages b and c are backlinks to page d

HITS is a link-based system that ranks webpages acquired from the Internet based on their textual content in response to a query. This function has certain drawbacks, as HITS usually provides more broad web pages on a topic that is normally quite specific. Overhead weight values are assigned to certain popular sites that aren't very relevant to the provided query.

The CLEVER algorithm is an extension of conventional HITS that gives a suitable answer to the issues that regular HITS might cause. CLEVER gives each link a weight depending on the terms of the searches and the link's endpoints.

It also uses anchor text to assign weights to the links. Large hub pages are broken down into smaller sections, allowing each hub page to focus on a specific topic.

PHITS is a probabilistic term-document relationship interpreter that recognises authoritative documents. The capacity of the PHITS algorithm to estimate real probabilities of authorities in comparison to the scalar magnitudes of authority supplied by traditional HITS is its most essential feature. In a comparison of results produced by conventional HITS versus PHITS on a set of hyperlinked papers, PHITS produces superior results.

4. The PageRank Algorithm

One of the most frequently used page ranking algorithms, the PageRank algorithm, considers backlinks and propagates the ranking through connections. If the total of a page's backlinks' ranks is high, it has a high rank.

4.1. Simplified PageRank

PageRank is defined in a somewhat simpler form as

$$PageRank(a) = k \sum_{b \in B(a)} \frac{PageRank(b)}{N_b}$$

where,

1) a is a web page
2) B(a) is a collection of web pages that point to a
3) PageRank(a) and PageRank(b) are the page a and b rank scores, respectively
4) N_b is the total number of outgoing links from page b
5) k is a normalisation factor

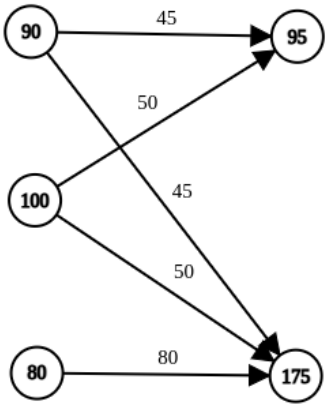

Figure 4. Simplified PageRank with normalisation factor k = 1

The rank score of a page is evenly distributed across its outbound links in PageRank. The rankings of the pages to which page p is directing are calculated using the values assigned to the links on page p. Two or more pages on a website may be linked together to form a loop. This is referred to as a rank sink scenario.

4.2. PageRank

Some people, for example, move straight to page b after seeing page a, even if page a is not directly connected to page b. Even though the two pages are not directly related, page a should impact the rank of page b in this situation. As a result, there is no such thing as an absolute rank sink.

The original PageRank is released after taking into account the phenomena stated above.

$$PageRank(a) = (1-d) + d \sum_{b \in B(a)} \frac{PageRank(b)}{N_b}$$

where,

1) d is the likelihood of people following the links
2) (1 - d) represents PageRank distribution from pages which are non-directly linked

Since the rank value converges in approximately log(n) of a site's popularity in the trials, the method performs efficiently and effectively.

Some connections on a web page may be more significant than others in the real world, resulting in a hierarchy that influences how a page ranks.

4.3. Weighted PageRank (WPR)

Weighted PageRank Algorithm (WPR) assigns greater rank values to more popular pages, instead of evenly distributing the rank value to its outlinks. The more popular a webpage is, the more links to it that other web pages have to it or are linked to it.

Different rank values are assigned to each of its outlinks based on their popularity.

Let's define the populatrity metrices of the link(b, a),

1) $W^{in}_{(b,\,a)}$, based on the no. of inlinks
2) $W^{out}_{(b,\,a)}$, based on the no. of outlinks

Now, some other important terms are,

1) No. of inlinks on page a, I_a
2) No. of inlinks on page p, I_p
3) No. of outlinks from page a, O_a
4) No. of outlinks from page p, O_p
5) Reference page list of page b, R(b)

$$W^{in}_{(b,\,a)} = \frac{I_a}{\sum_{p \in R(b)} I_p}$$

and

$$W^{out}_{(b,\,a)} = \frac{O_a}{\sum_{p \in R(b)} O_p}$$

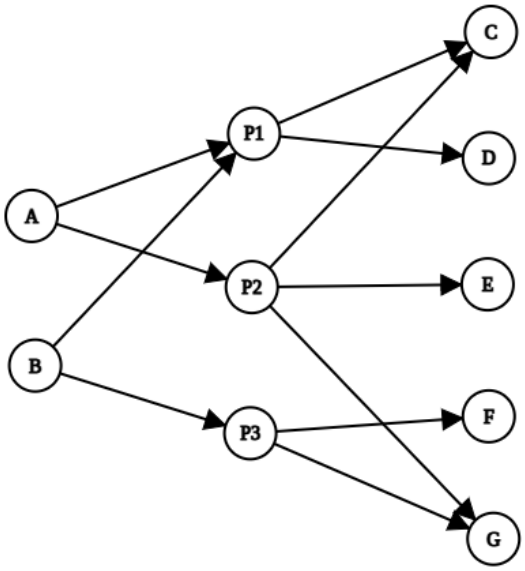

Figure 5. Weighted PageRank (WPR) example

Page A has two reference pages : p1 and p2. I_{p1} = 2, I_{p2} = 1, O_{p1} = 2, and O_{p2} = 3 are the inlinks and outlinks of these two pages.

$$W^{in}_{(A,p1)} = \frac{I_{p1}}{I_{p1} + I_{p2}} = \frac{2}{2+1} = \frac{2}{3}$$

and

$$W^{\text{out}}_{(\text{A,p1})} = \frac{O_{p1}}{O_{p1} + O_{p2}} = \frac{2}{2+3} = \frac{2}{5}$$

This implies the updated PageRank formula is,

$$PageRank(a) = (1-d)+d \sum_{b \in B(a)} PageRank(b) W^{\text{in}}_{(\text{b, a})} W^{\text{out}}_{(\text{b, a})}$$

5. Conclusion

Web mining is a technique for extracting data from user activity, and web structure mining is a key component. HITS and PageRank are two frequently used algorithms for ranking the relevant sites. When allocating rank scores, both algorithms consider all connections equally.

This work introduces the WPR algorithm, which is a PageRank extension. WPR awards rank points to pages based on their popularity, taking into account the value of both inward and outbound connections. In the current version of WPR, just the inlinks and outlinks of the pages in the reference page list are used to generate the rank scores. We'd want to look at the potential of employing several levels of reference page listings in our future research on this approach.

Acknowledgments

Dr. Arti Kashyap has been really generous in allowing me to participate in this study, which has provided me with vital information regarding Google's Pagerank and web search.

6. References

1) The PageRank Citation Ranking:Bringing Order to the Web
2) PageRank: Standing on the shoulders of giants
3) The Google Pagerank Algorithm and How It Works