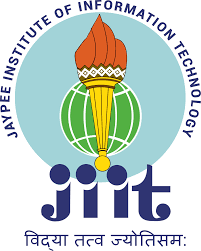
JAYPEE INSTITUTE OF INFORMATION

TECHNOLOGY



     TOPIC: WEB PAGE RANK ALGORITHM

Submitted By: Submitted To:

Rachit Goel (17103146) DR. Pawan Kr. Upadhyay

Shagun Agarwal (17803003) Mr. Prantik Biswas

Astitva Singhal (17103136)

Ankit Vidyarthi (Faculty Supervisor)

**CONTENT**

1. INTRODUCTION
2. TECHNOLOGY USED
3. EXISTING ALGORITHMS
4. RESULTS
5. FUTURE WORK
6. REFERENCES

**INTRODUCTION**

PageRank is a topic widely discussed by Search Engine Optimization (SEO) experts. The heart of PageRank is a mathematical formula which seems to be very complicated, but it is actually simple to understand if simplified. It is necessary for the web developers to understand the concept of PageRank to make the page more high ranked. Since automation is making the life of a person easier, automating PageRank will have a better effect on the World Wide Web. It was invented by Larry Page and Sergey Brin while they were graduate students at Stanford University. There is a math behind every algorithm. Matrix and vector can be considered as the main source of many achievements.

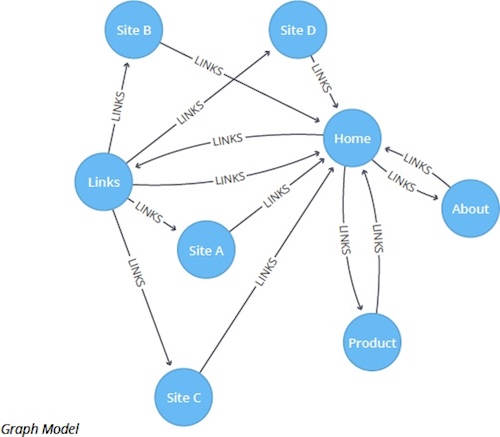
**Page Rank works by counting the numbers and quality of inlinks and outlinks to a page to determine a rough estimate of how important the page is.**

Usually, the user expects the relevant page to be displayed in the top 20-30 pages provided by search engine. A modern search engine uses the method of providing best results first that are more appropriate than the older text ranking method. One of the most influential algorithm is the Page Rank algorithm. The main idea behind the page rank algorithm is that the importance of a web page is predicted by the pages linking to it. If we create a web page i that has connected page j then page j is considered as important. On the other hand, if page j has a backlink from page k we can say k transfers its control to j (i.e., k asserts that j is important). We can iteratively assign a rank to each page based on the number of pages that points to it.

For Eg:

The following Cypher statement creates a sample graph of web pages and links between

them.



**Problem Statement:**

In modern world, websites are growing bigger by the day. Bigger a website more the number of web pages and content. When a user searches for information within the website, he’s get a whole set of pages containing parts of his search phrase and has to browse through them all to find relevant information. We want to use the power of page ranking algorithms to develop efficient ranking systems so that the best result is showed first.

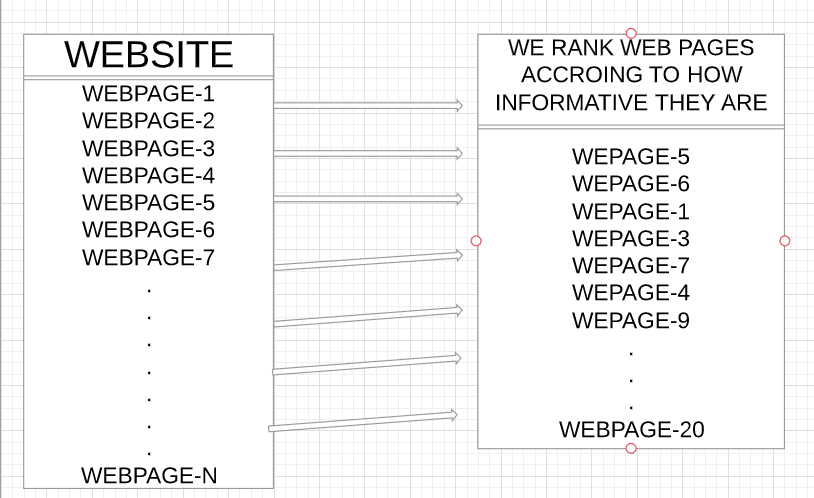
**Technologies Used:**

**EmEditor**

**Spyder (Python IDLE)**

**Languages:**

**Python3**

****

Algorithms We Implemented:

1. Page Rank Algorithms:

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element E is referred to as the PageRank of E and denoted by **PR(E).**

The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high PageRank receives a high rank itself.

PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to more closely reflect th theoretical true value. A probability is expressed as a numeric value between 0 and 1. A 0.5 probability is commonly expressed as a "50% chance" of something happening. Hence, a PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank.

Algorithm:

Suppose instead that page **B** had a link to pages **C** and **A**, page **C** had a link to page **A**, and page **D** had links to all

three pages. Thus, upon the next iteration, page **B** would transfer half of its existing value, or 0.125, to page **A** and the other half, or 0.125, to page **C**. Page **C** would transfer all of its existing value, 0.25, to the only page it links to,

**A**. Since **D** had three outbound links, it would transfer one third of its existing value, or approximately 0.083, to **A**.

At the completion of this iteration, page **A** will have a PageRank of 0.458.





In other words, the PageRank conferred by an outbound link is equal to the document's own PageRank score divided by the number of outbound links **L( )**.



In the general case, the PageRank value for any page **u** can be expressed as:



i.e. the PageRank value for a page **u** is dependent on the PageRank values for each page **v** contained in the set **Bu**

(the set containing all pages linking to page **u**), divided by the number *L*(*v*) of links from page **v**.

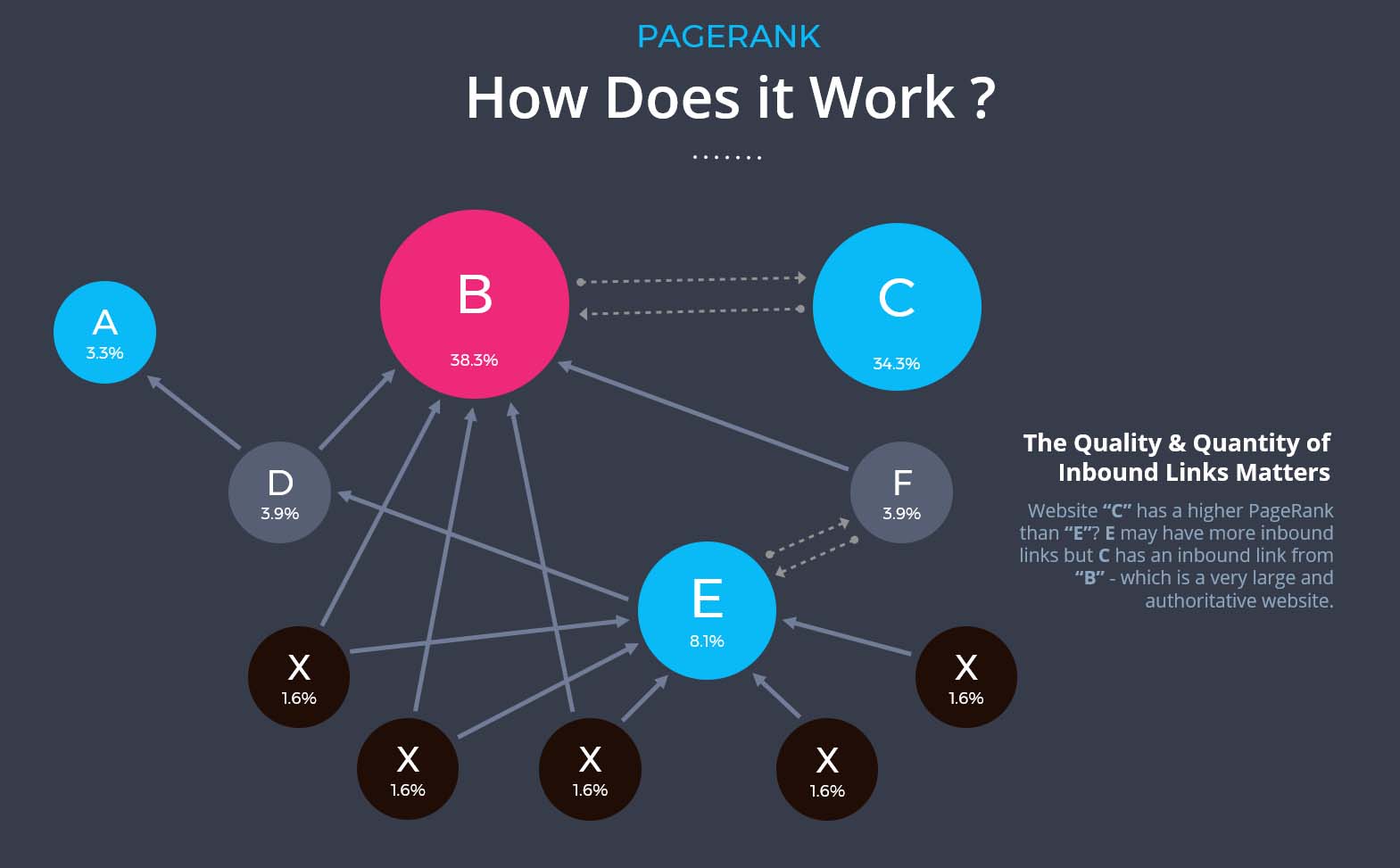
**Damping factor:**

The PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking.

The probability, at any step, that the person will continue is a damping factor *d*. Various studies have tested different damping factors, but it is generally assumed that the damping factor will be set around 0.85.



where  are the pages under consideration,  is the set of pages that link to  ,  is the number of outbound links on page pj, and *N* is the total number of pages. The PageRank values are the entries of the dominant eigenvector of the modified adjacency matrix. This makes PageRank a particularly elegant metric.

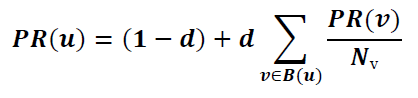


1. Weighted Page Rank Algorithm:

In this paper, a page ranking mechanism called Weighted PageRank Algorithm based on Visits of Links (VOL) is being devised for search engines, which works on the basis of Weighted PageRank Algorithm and takes number of visits of inbound links of web pages into account. The original Weighted PageRank algorithm (WPR) is an extension to the standard PageRank algorithm. WPR takes into account the importance of both the inlinks and outlinks of the pages and distributes rank scores based on the popularity of the pages. The main purpose of the proposed algorithm is finding more relevant information according to user’s query. So, this concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behaviour, which reduce the search space to a large scale.

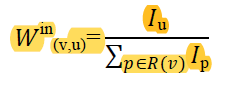
Algorithm:

Where u represents a web page, B (u) is the set of pages that point to u, PR (u) and PR (v) are rank scores of page u and v respectively, Nv indicates the number of outgoing links of page v, c is a factor applied for normalization. Later PageRank was customized observing that not all users follow the direct links on WWW. The modified version is given in Eq. 2:

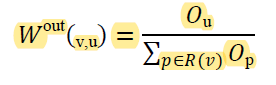


a new approach known as weighted pagerank algorithm (WPR). This algorithm is an extension of PageRank algorithm. WPR takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages. WPR performs better than the conventional PageRank algorithm in terms of returning larger number of relevant pages to a given query.

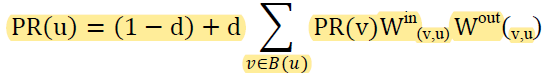
The proposed extended PageRank algorithm–a Weighted PageRank Algorithm–assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as *Win*(*v,u*) and *Wout*(*v,u*) , respectively. *Win*(*v,u*) given in eq. (3) is the weight of *link*(*v, u*) calculated based on the number of inlinks of page *u* and the number of inlinks of all reference pages of page *v*.



Where *Iu* and *Ip* represent the number of inlinks of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*. *Wout*(*v,u*) given in eq. (4) is the weight of *link*(*v, u*) calculated based on the number of outlinks of page *u* and the number of outlinks of all reference pages of page *v.*



Considering the importance of pages, the original PageRank formula is modified as



3) Hyperlink Topic Induced Search

Hypertext Induced Topic Search (HITS) or hubs and authorities is a link analysis algorithm .A precursor to PageRank ,HITS is a search query dependent algorithm that ranks the web pages by processing its entire in links and out links .Thus ,ranking of the web pages 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 :

**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.

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 page 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.

Figure below depicts the hubs and authorities created by HITS.

Authorities and hubs exhibit a mutually reinforcing relationship : a better hub points to many good authorities, and a better authority is pointed to by many good hubs.

To mark a web page as Authority or Update, HITS follows the following rules :

Authority Update Rule: ∀p, update auth (p) as follows:

Σi=1 hub(i) (1)

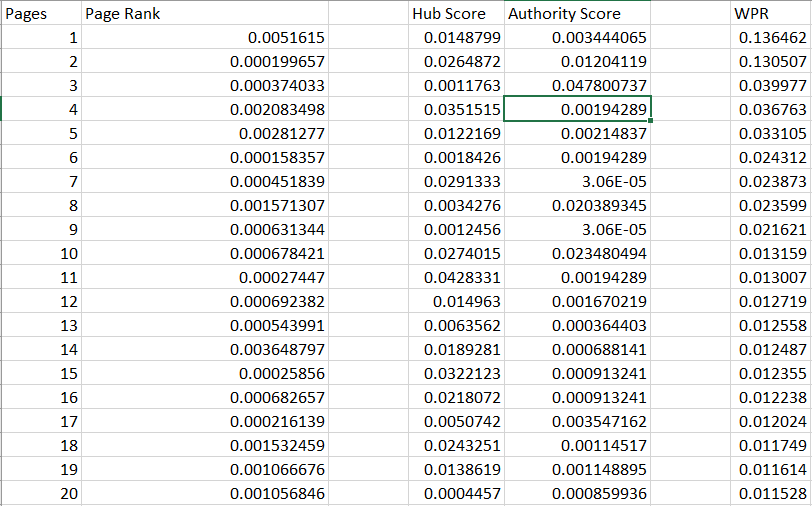
Where n is the total number of pages connected to p. According to (1) the Authority score of a page is the sum of all the Hub scores of pages that point to it .

Hub Update Rule: ∀p, we update hub (p) as follows:

𝑎𝑢𝑡ℎ Σ 𝑖=1 auth(i) (2)

Where n is the total number of pages, p connects to. According to (2) a page's Hub score is the sum of the Authority scores of all its linking pages . More precisely, given a set of web pages (say, retrieved in response to a search query), the HITS algorithm first forms the n by n adjacency matrix A, whose m( i , j) element is 1 if page i links to page j and 0 otherwise.

**RESULTS OF THE ALGORITHMS IMPLEMENTED:**



**FUTURE WORK:**

1. We hope to develop our own algorithm for ranking these pages with new parameters to judge relevancy of a page and less complexity than existing algorithms
2. We will perform precision and recall tests on the algorithms we have implemented and check if our algorithm provides better result than existing algorithms
3. We will demonstrate our algorithm in a real world scenario and check if its quality matches with the simulations

**REFERENCES**:

1. S. Prabha, K. Duraiswamy, J. Indhumathi, “Comparative Analysis of Different Page Ranking Algorithms”, International Journal of Computer and Information Engineering, Vol:8, No:8, 2014.
2. N. Tyagi, S. Sharma, “Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page”, International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-3, July 2012.
3. P.Sharma, D.Tyagi, P.Bhadana,” Weighted Page Content Rank for Ordering Web Search Result” International Journal of Engineering Science and Technology, 7301-7310, 2010.
4. Liviu P. Dinu and Radu Tudor Ionescu, “An Efficient Algorithm for Rank Distance”, Faculty of Mathematics and Computer Science University of Bucharest, December 2013.
5. P.Patel, K.Patel,” A Review of PageRank and HITS Algorithms”, International Journal of Advance Research in Engineering, Science & Technology(IJAREST), Volume 2,Issue 1, January- 2015
6. Hema Dubey ,Prof. B. N. Roy,” An Improved Page Rank Algorithm”, Department of Computer Science and Engineering, Maulana Azad National Institute of technology, Bhopal, International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2011, 2183-2188