**PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE**

**ACADEMIC YEAR: 2023-24**

## **DEPARTMENT of COMPUTER ENGINEERING DEPARTMENT**

**CLASS: B.E. SEMESTER: I**

**SUBJECT: LP-IV**

|  |  |
| --- | --- |
| **ASSIGNMENT NO.** | A2 |
| **TITLE** | **Page Rank Algorithm** |
| **PROBLEM STATEMENT /DEFINITION** | Implement **Page Rank Algorithm**. |
| **OBJECTIVE** | To perform ranked retrieval of documents by assigning page ranks to each document as per the importance of the document. |
| **OUTCOME** | Students will be able to understand:   * Perform ranked retrieval of the documents according to their importance by assigning them page ranks. * How Page Rank Algorithm works and used in various search engines. |
| **S/W PACKAGES AND**  **HARDWARE APPARATUS USED** | Python 3.9.0,  VS Code,  Windows (64-bit),  Intel I5 4GB RAM |
| **REFERENCES** | 1. C.J. Rijsbergen, &quot;Information Retrieval&quot;, (http://www.dcs.gla.ac.uk/Keith/Preface.html)  2. W.R. Hersh, ―Information Retrieval: A Health and Biomedical Perspective‖, Springer, 2002.  3. G. Kowalski, M.T. Maybury. &quot;Information storage and Retrieval System&quot; , Springer, 2005 |
| **STEPS** | Refer to theory, algorithm, test input, test output |
| **INSTRUCTIONS FOR**  **WRITING JOURNAL** | 1. Date  2. Assignment no.  3. Problem definition  4. Learning objective  5. Learning Outcome  6. Concepts related Theory  7. Algorithm  8. Test cases  10. Conclusion/Analysis |

**Prerequisites:**

**Concepts related Theory:**

**Page Rank Algorithm**

PageRank (PR) is an algorithm used by Google Search to rank web pages in their search engine results. It is named after both the term "web page" and co-founder Larry Page. PageRank is a way of measuring the importance of website pages.

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

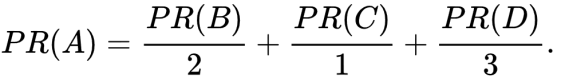
PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal 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).

Assume a small universe of four web pages: **A**, **B**, **C**, and **D**. Links from a page to itself are ignored. Multiple outbound links from one page to another page are treated as a single link. The initial value of the web pages can be assumed to be anything between 0 and 1, let us say 0.25.

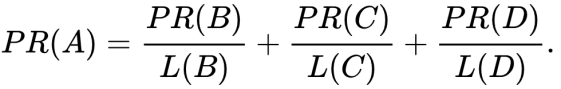
If the only links in the system were from pages **B**, **C**, and **D** to **A**, each link would transfer 0.25 PageRank to **A** upon the next iteration, for a total of 0.75.



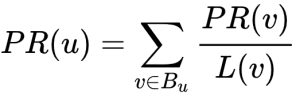
Suppose instead that page **B** had a link to pages **C** and **A**, page **C** had a link to page **A**, and page **D** had links to all three pages. Thus, upon the first iteration, page **B** would transfer half of its existing value (0.125) to page **A** and the other half (0.125) to page **C**. Page **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 approximately 0.458.



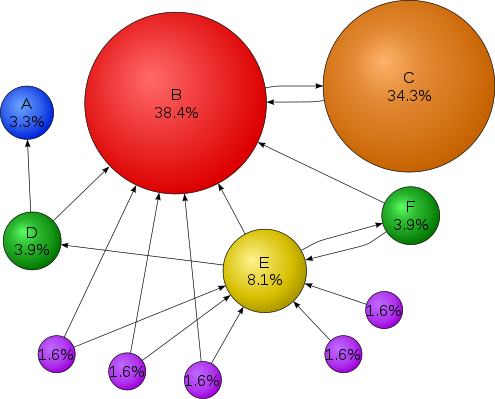
In other words, the PageRank conferred by an outbound link is equal to the document's own PageRank score divided by the number of outbound links **L( )**.



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**.



Above is the simple illustration of the Page Rank algorithm. The percentage shows the perceived importance, and the arrows represent hyperlinks.

**Algorithm:**

Iteration 0: Initialize all ranks to be 1/(number of total pages).

Iteration 1: For each page u, update u’s rank to be the sum of each incoming page v’s rank from the previous iteration, divided by the number total number of links from page v.

**Conclusion:** Page Rank Algorithm is implemented successfully using python programming language for ranked retrieval of documents.

**Review Questions**:

Q1. What are page speed and its role in SEO?

Q2. What are meta tags? Name the important ones and their character limits.

Q3. What is a landing page?

Q4. Explain – spiders, crawlers and robots.