Title: Implementation of PageRank & HITS algorithms.

Objective/Aim: To perform the following tasks:

1. Implement the PageRank algorithm to calculate the rank of each page in the file. The output should be the 10 pages with the highest rank, together with their rank values.
2. Implement the HITS algorithm to calculate the hub and the authority weight of each web page in the data set. The output should be the 10 most authoritative pages and 10 most hubby pages.
3. Tabulate the results containing adjacency matrix and rank of pages.

Introduction: This assignment is about implementation of algorithms used by search engines for ranking and searching of webpages in order to retrieve more relevant information.

Theory/Algorithm:

1. Pagerank algorithm: 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.

   How does it work ?
   The PageRank algorithm outputs 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 the theoretical true value.

   Simplified algorithm
   Assume a small universe of four web pages: A, B, C, and D. Links from a page to itself, or multiple outbound links from one single page to another single page, are ignored. PageRank is initialized to the same value for all pages. In the original form of PageRank, the sum of PageRank over all pages was the total number of pages on the web at that time, so each page in this example would have an initial value of 1. However, later versions of PageRank, and the remainder of this section, assume a probability distribution between 0 and 1. Hence the initial value for each page in this example is 0.25.

Authors: 2018BTECS00050
2018BTECS00064

The PageRank transferred from a given page to the targets of its outbound links upon the next iteration is divided equally among all outbound links. 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.
PR(A) = PR(B) + PR(C) + PR(D).\,
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, 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 approximately 0.458.
PR(A)={\frac {PR(B)}{2}}+{\frac {PR(C)}{1}}+{\frac {PR(D)}{3}}.\,
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( ).
PR(A)={\frac {PR(B)}{L(B)}}+{\frac {PR(C)}{L(C)}}+{\frac {PR(D)}{L(D)}}.\, In the general case, the PageRank value for any page u can be expressed as:
PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} ,
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. The algorithm involves a damping factor for the calculation of the PageRank. It is like the income tax which the govt extracts from one despite paying him itself.

2. HITS algorithm: Hyperlink Induced Topic Search (HITS) Algorithm is a Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg. This algorithm is used to the web link-structures to discover and rank the webpages relevant for a particular search.
HITS uses hubs and authorities to define a recursive relationship between webpages. Before understanding the HITS Algorithm, we first need to know about Hubs and Authorities.

   Given a query to a Search Engine, the set of highly relevant web pages are called Roots. They are potential Authorities.
Pages that are not very relevant but point to pages in the Root are called Hubs. Thus, an Authority is a page that many hubs link to whereas a Hub is a page that links to many authorities.

Authors: 2018BTECS00050
2018BTECS00064

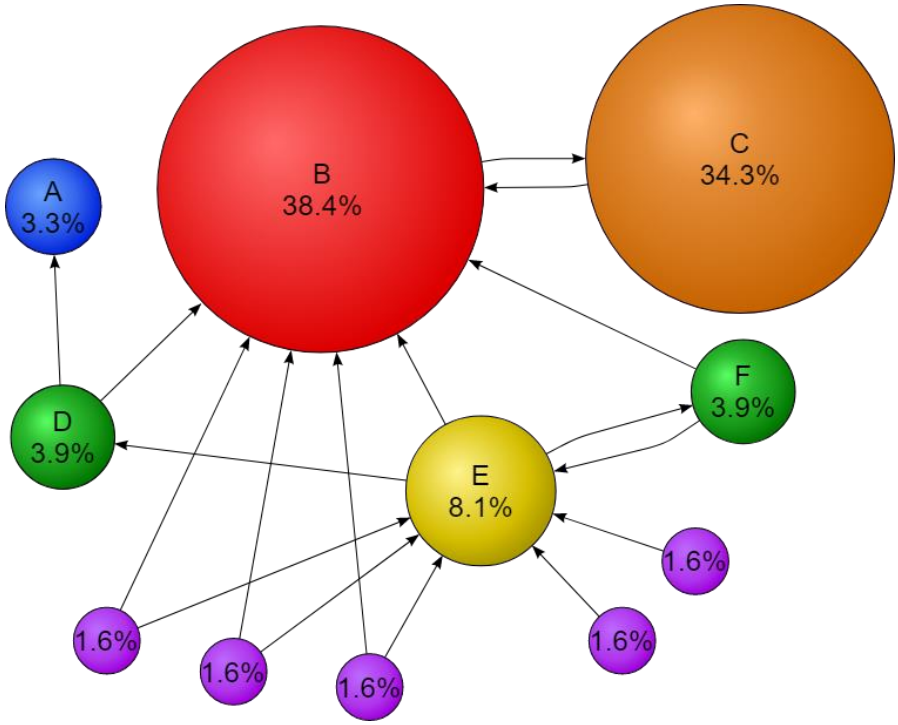
Pseudocode:

```
1 Let G be set of pages
2 for each page pg in G do
3 pg.auth = 1 // authority score of the page pg
4 pg.hub = 1 // hub score of the page pg
5 function Calc_Hubs_Authorities(G)
6 for step from 1 to i do // run the algorithm for i steps
7 norm = 0
8 for each page pg in G do // update authority values
9 pg.auth = 0
10 for each page qg in p.inNeighbors do //set of pages that link to pg
11 pg.auth += qg.hub
12 norm += square(pg.auth) //sum of the squared auth values to normalise
13 norm = sqrt(normal)
14 for each page pg in G do // update the auth scores
15 pg.auth = pg.auth / normal // normalise the auth values
16 norm = 0
17 for each page pg in G do // update hub values
18 pg.hub = 0
19 for each page rg in pg.outNeighbors do // set of pages that pg links to
20 pg.hub += rg.auth
21 norm += square(pg.hub) //sum of the squared hub values to normalise
22 norm = sqrt(normal)
23 for each page pg in G do //update hub values
24 pg.hub = pg.hub / normal // normalise the hub values
```

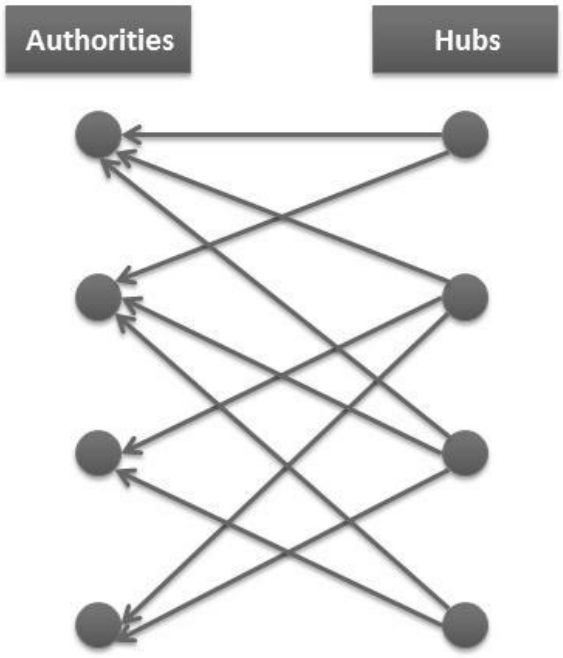
Documentation/Block Diagrams:

Pagerank Algorithm:

Authors: 2018BTECS00050
2018BTECS00064



HITS algorithm:



Conclusion:

From this assignment we understood the underlaying logic of a web search engine.

1. Pagerank algorithm:

Authors: 2018BTECS00050
2018BTECS00064

a. Advantages:
    i. Since it pre computes the rank score it takes less time and hence it is fast.
    ii. It is more feasible as it computes rank score at indexing time not at query time
    iii. It returns important pages as Rank is calculated on the basis of the popularity of a page.
b. Disadvantages:
    1. 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.
    2. Relevancy of the resultant pages to the user query is very less as it does not consider the content of web page.
    3. 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.
    4. Dead Ends are possible ie., pages with no outgoing links.
    5. 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.
    6. If you have circle references in your website, then it will reduce your front page's PageRank.

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

i. Query Time cost: The query time evaluation is expensive. This is a major drawback since HITS is a query dependent algorithm.
ii. Irrelevant authorities: 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).
iii. Irrelevant Hubs: 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.
iv. Mutually reinforcing relationships between hosts: HITS emphasizes mutual reinforcement between authority and hub webpages. 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.
v. Topic Drift: 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.
vi. Less Feasibility: 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). Because this computation is carried out at query time, it is not feasible for today's search engines, which need to handle tens of millions of queries per day

References:

1. https://www.ijert.org/research/comparative-analysis-of-pagerank-and-hits-algorithms-IJERTV1IS8530.pdf
2. https://www.deepcrawl.com/knowledge/technical-seo-library/how-do-search-engines-

work/#:~:text=Search%20engines%20work%20by%20crawling,that%20have%20been%20made%20available.
3. https://en.wikipedia.org/wiki/PageRank

Authors: 2018BTECS00050
2018BTECS00064