**Assignment 7**

**Group: DWDM19G04**

**Roll Numbers: 2016BTECS00063, 2016BTECS00081, 2016BTECS00103**

**Batch: B7**

**Date: 4th November, 2019**

**Title: Web Mining**

**Problem Statement:**

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.

• Use Web Graph datasets from “Stanford Large Network Dataset Collection” http://snap.stanford.edu/data/#web

• Tabulate the results.

**Theory:**

**1. Page Rank Algorithm**

PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. 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. The underlying assumption is that more important websites are likely to receive more links from other websites.

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.

The above centrality measure is not implemented for multi-graphs.

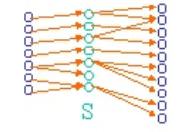
**2. HITS algorithm:**

**Hyperlink-Induced Topic Search (HITS)** (also known as Hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It determines two values for a page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

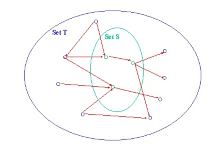
**Steps involved in HITS algorithm:**

1. Starting from the user supplied query, HITS assembles initial set S of pages:

The initial set of pages is called root set. These pages are then expanded to a larger root set T by adding any pages that are linked to or from any page in the initial set S.



1. HITS then associates with each page p a hub weight h(p) and an authority weight a(p), all initialized to 1.



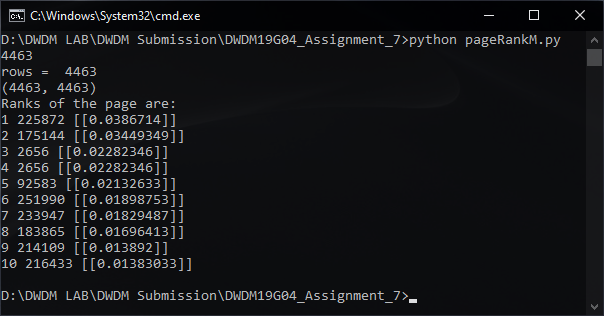
1. HITS then iteratively update the hubs and authority weights of each page. Let p → q denotes “page p has an hyperlink to page q”. HITS updates the hubs and authority as follows:

a(p)=∑p→qh(q)a(p)=∑p→qh(q)

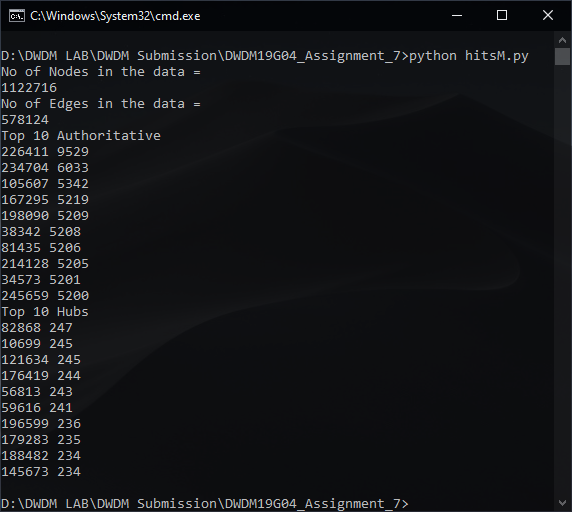
h(p)=∑q→pa(q)

**Output:**

**Page Rank Algorithm:**

****

**HITS Algorithm:**

****

**Tabular Result:**

**Top 10 Authoritative:**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Page ID** | **Authority Score** |
|  |  |  |
| **1** | 226411 | 9529 |
| **2** | 234704 | 6033 |
| **3** | 105607 | 5342 |
| **4** | 167295 | 5219 |
| **5** | 198090 | 5209 |
| **6** | 38342 | 5208 |
| **7** | 81435 | 5206 |
| **8** | 214128 | 5205 |
| **9** | 34573 | 5201 |
| **10** | 245659 | 5200 |

**Top 10 Hubs:**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Page ID** | **Hub Score** |
| **1** | 82868 | 247 |
| **2** | 10699 | 245 |
| **3** | 121634 | 245 |
| **4** | 176419 | 244 |
| **5** | 56813 | 243 |
| **6** | 59616 | 241 |
| **7** | 196599 | 236 |
| **8** | 179283 | 235 |
| **9** | 266637 | 234 |
| **10** | 145673 | 234 |

**Conclusion:**

HITS uses hubs and authorities to define a recursive relationship between webpages. 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.