# Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

**AY: 2025-26**

| | | | |
|---|---|---|---|
| **Class:** | TE | **Semester:** | V th |
| **Course Code:** | CSC 504 | **Course Name:** | Data warehousing & mining |

| | |
|---|---|
| **Name of Student:** | Pranita kumbhar |
| **Roll No. :** | 70 |
| **Assignment No.:** | 06 |
| **Title of Assignment:** | Web mining. |
| **Date of Submission:** | |
| **Date of Correction:** | |

## Evaluation

| Performance Indicator | Max. Marks | Marks Obtained |
|---|---|---|
| Completeness | 5 | 5 |
| Demonstrated Knowledge | 3 | 3 |
| Legibility | 2 | 2 |
| Total | 10 | 10 |

| Performance Indicator | Exceed Expectations (EE) | Meet Expectations (ME) | Below Expectations (BE) |
|---|---|---|---|
| Completeness | 5 | 3-4 | 1-2 |
| Demonstrated Knowledge Legibility | 3 | 2 | 1 |
| Legibility | 2 | 1 | 0 |

## Checked by

**Name of Faculty** : Ms. Neha Raut

**Signature** :

**Date** :

Q.1] Explain in detail with its significance in ranking web pages and how it models the importance of a page.

→

- **PageRank Algorithm:**

- PageRank is an algorithm developed by Larry Page and Sergey Brin, the founders of Google, to measure the founders of Google, to measure the importance of web pages. It assigns a numerical weight to each web page to determine its relative importance based on the number & quality of links pointing to it.

- **Concept:**
- Each link to a web page is considered a 'vote' for that page.
- However, votes from important pages carry more weight.
- The page rank value of a page depends on the pagebank of the pages linking to it.

Formula:

$$PR(A) = (1-d) + d\left(\sum \frac{PR(T_i)}{C(T)_i}\right)$$

- $PR(A)$ = PageRank of page A.
- $d$ = damping factor (usually 0.85)
- $T_i$ = pages linking to A.
- $C(T_i)$ = number of outgoing links from page $T_i$.

- **Significance in Ranking:**
- Helps search engines rank web pages based on relevance and popularity.

- Pages with more high-quality inbound links rank higher.
- Reduces the effect of spam or low-quality links.

- How it Models Importance :

- Models the web as a directed graph.
- A page's importance is determined recursively by the importance of the pages linking to it.
- Reflects "real-world" popularity : a page linked by many important pages becomes important itself.

- Example :

If Page A is linked by highly ranked pages B and C, Page A will have a higher PageRank than a page linked only by less popular pages.

Output :

PageRank assigns a numerical score (e.g. 0 to 1) to each page, used to order results in search engine rankings.

Q.2] Explain the in detail. Define hub and authority and discuss their conceptual roles in web page ranking.

→

- HITS Algorithm (Hyperlink - Induced Topic Search):
HITS is an algorithm proposed by Jon Kleinberg that ranks web pages based on two scores: Authority and Hub scores. It helps identify relevant pages for a particular topic.

- Concept :
  - Authority Pages : Pages that contain valuable information about a topic.
  - Hub pages : Pages that link to many authority pages.

Each page has both a hub score and an authority score, and these values reinforce each other.

- Working steps :

1] Construct a root set of relevant web pages based on query.
2] Expand it to include pages that link to or are linked from the root set.
3] Iteratively update authority and hub scores :
  - Authority score of a page = sum of hub scores of pages linking to it.
  - Hub score of a page = sum of authority scores of pages it links to.
4] Normalize the scores until they converge.

- **Formula :**

- $A(p) = \Sigma_{q \to p} H(q)$
- $H(p) = \Sigma_{p \to r} A(r)$

Where,

- $A(p)$ = authority score of page $p$.
- $H(p)$ = hub score of page $p$.
- $q \to p$ = means page $q$ links to $p$.
- $p \to r$ = means page $p$ links to $r$.

- **Conceptual Roles :**

- **Hub** - Acts like a directory - links to many useful authority pages.
- **Authority** - Acts ~~link~~ like a trusted source → receives links from many hub pages.

- **Example** - If Pages X links to Pages Y & Z (both good authorities), Page X becomes a strong hub. If many hub pages link to Page Y, Page Y becomes a authority.

- **Significance in web page ranking** -

- HITS is used to identify topic-specific authoritative page.
- It improves relevance in focused searches.
- Unlike PageRank, it distinguishes between hubs & authority.

- **Output :**

Each web page gets two scores - Authority Score & Hub Score, - which helps rank pages in topic-based searches.