NAME :______ STD.:______ DIV.:______

NIRZARI PARIKH
60004210156
BATCH-C22

WEB INTELLIGENCE (WI)

ASSIGNMENT 01

Q.1 Explain web spamming. Describe in detail the different types of web spamming. Provide solution for web spamming.

① Web spamming refers to misleading activities aimed at boosting a webpage's ranking in search results without increasing its actual information value.

② Spammers exploit weakness in search engine algorithms to manipulate rankings, leading to poor search experiences.

Types of web spamming

① Content spamming

a] Manipulating text fields (title, mega-tags, body, anchor text URL) to increase keyword relevance

b] Techniques include excessive keyword repetition and adding unrelated popular terms to attract traffic.

② Link spamming

a] Out-link spamming - Adding many links to authoritative sites to boost credibility

b] In-link spamming - Gaining backlinks through honeypots, directory submissions, forums / blog posts, link exchanges or creating spam farms.

③ Hiding Techniques

a] Content Hiding - Using same-color text as the background or hidden elements.

b] Cloaking - Showing different content to users and search engines.

c] Redirection - Automatically redirecting users from a spammed page to a different one.

NAME : STD.: DIV.:

NIRZARI PARIKH
60004210156

Solutions to Web Spamming.

① Advanced Search Algorithm - Improve spam detection by analyzing content patterns.

② Penalizing spam Pages - Reduce ranking or remove flagged spam pages.

③ User Feedback - Allow reporting of suspicious websites.

④ Machine learning & AI - Automate spam detection using intelligent algorithms.

⑤ Regular Index Updates - Ensure search engines refresh rankings frequently to filter out spam.

.2 Explain vector Space Model clearly with an example.

① The Vector Space Model (VSM) is a popular information retrieval model where documents and queries are represented as vectors in multi-dimensional space.

② The relevance of a document to a query is determined by measuring the similarity between the vectors.

③ Document Representation.

a] Each document is a vector with term weights based on TF (Term Frequency) or TF-IDF (Term Frequency - Inverse Document Frequency)

b] Unlike Boolean models, term weights are real values, not just 0 or 1.

④ TF - IDF Weighting

a] TF - ~~Term~~ Frequency of a term in a document

b] IDF - Reduces weight of common terms.

c] Formula

$w_{ij} = tf_{ij} \times idf_{i1}$, where $idf_{i1} = \log(N/df_i)$

NIRZARI PARIKH

60004210156

⑤ Similarity Measurement (cosine similarity)

$$\cos(q, d_j) = \frac{\sum w_{iq} \, w_{ij}}{\sqrt{\sum w_{iq}^2} \, \sqrt{\sum w_{ij}^2}}$$

⑥ Example

Query : "machine learning"

a] Documents :

D1 : " Machine Learning is powerful."

D2 : " Deep Learning is a subset of machine learning!"

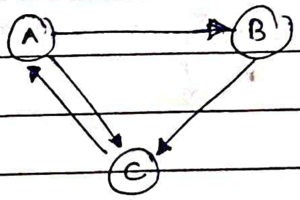
b] TF - IDF weights assigned, and cosine similarity ranks documents based on relevance.

⑦ Advantages

a] Ranks documents by relevance, not strict matching

b] Handles partial matches, widely used in search engines.

Q.3



d = 0.85

Initial Page Rank = 1

Iteration = 3.

Iteration ①

$$PR(A) = (1-d) + d \left[\frac{PR(C)}{1} \right] = (1-0.85) + 0.85 \left[\frac{1}{1} \right]$$

PR(A) = 1

$$PR(B) = (1-d) + d \left[\frac{PR(A)}{2} \right] = (1-0.85) + 0.85 \left[\frac{1}{2} \right]$$

PR(B) = 0.575

NIRZARI PARIKH

60004210156

$$PR(C) = (1-d) + d\left[\frac{PR(A)}{2} + \frac{PR(B)}{1}\right]$$

$$PR(C) = (1-0.85) + 0.85\left[\frac{1}{2} + \frac{0.575}{1}\right]$$

$$PR(C) = 1.064$$

Iteration (2) $PR(A) = (1-d) + d\left[\frac{PR(C)}{1}\right] = (1-0.85) + 0.85\left[\frac{1.064}{1}\right]$

$$PR(A) = 1.054$$

$$PR(B) = (1-d) + d\left[\frac{PR(A)}{2}\right] = (1-0.85) + 0.85\left[\frac{1.054}{2}\right]$$

$$PR(B) = 0.598$$

$$PR(C) = (1-d) + d\left[\frac{PR(A)}{2} + \frac{PR(B)}{1}\right] = (1-0.85) + 0.85\left[\frac{1.054}{2} + \frac{0.598}{1}\right]$$

$$PR(C) = 1.106$$

Iteration (3) $PR(A) = (1-d) + d\left[\frac{PR(C)}{1}\right] = (1-0.85) + 0.85\left[\frac{1.106}{1}\right]$

$$PR(A) = 1.09$$

$$PR(B) = (1-d) + d\left[\frac{PR(A)}{2} \cancel{+ \frac{PR(B)}{1}}\right] = (1-0.85) + 0.85\left[\frac{1.09}{2}\right]$$

$$PR(B) = 0.613$$

$$PR(C) = (1-d) + d\left[\frac{PR(A)}{2} + \frac{PR(B)}{1}\right] = (1-0.85) + 0.85\left[\frac{1.09}{2} + 0.613\right]$$

$$PR(C) = 1.134$$

NIRZARI PARIKH
60004210156.

Q.4 List and explain implementation issues in web crawling.

① Scalability :- The web is vast and continuously growing. crawlers must be efficient in managing large-scale data while balancing storage and bandwidth constraints.

② Fetching & Parsing - crawlers need to efficiently download and parse web pages while handling different formats (HTML, XML, JSON) and errors like broken links or slow responses.

③ Stopword Removal & Stemming - To improve indexing, unnecessary words (stopwords are removed), and stemming reduces words to their root forms (eg: "running" → "run").

④ Link Extraction & Canonicalization
Extracting links ensure proper navigation, while canonicalization avoids duplicate content by normalizing URLs (eg: handling www vs. non-www versions).

⑤ Spider Traps - Dynamically generated links, infinite loops and session-based URLs can cause crawlers to get stuck, wasting resources.

⑥ Page Repository Management - Efficient storage and indexing of crawled data are essential, requiring deduplication and proper version control for updated content

⑦ content & Parallelization - crawlers must handle multiple requests simultaneously using multi-threading or distributed systems to speed up crawling without overloading servers.

Q.5 Illustrate HITS algorithm with an example.

① HITS (Hyperlink - Induced Topic Search) is a link analysis algorithm used to rank web pages based on their importance. It identifies two types of pages:

NIRZARI PARIKH
60004210156

a] Hubs : Pages with many outgoing links to important pages.

b] Authorities : Pages with many incoming links from good hubs.

② steps in HITS algorithm.

a] construct a link Graph – Represent webpages as nodes and hyperlinks as directed edges.

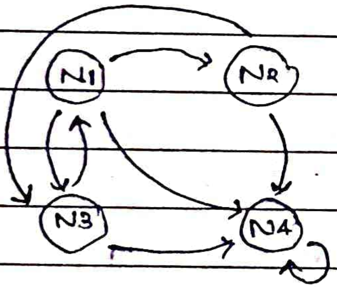
b] Initialize Hub & Authority Scores – Assign each node an initial score.

c] Update Authority scores – sum the hub scores of all linking pages

d] update Hub scores – sum the authority scores of all linked pages

e] Normalize scores – scale scores to prevent exponential growth.

f] Iterate until convergence – Repeat steps 3-5 until scores stabilize.



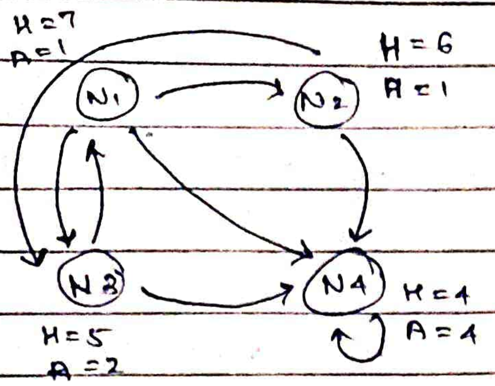
$$A = \begin{array}{c|cccc} & N1 & N2 & N3 & N4 \\ \hline N1 & 0 & 1 & 1 & 1 \\ N2 & 0 & 0 & 1 & 1 \\ N3 & 1 & 0 & 0 & 1 \\ N4 & 0 & 0 & 0 & 1 \end{array} \qquad A^T = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Iteration ① $V = A^T * u$ ⟶ Authority weight vector

$$V = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 4 \end{bmatrix}$$

$u = A * V$ ⟶ updated Hub weight vector

$$u = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 7 \\ 6 \\ 5 \\ 4 \end{bmatrix}$$



Nodes	Hub	Authority
N1	7	1
N2	6	1
N3	5	2
N4	4	4
	N1 N2 N3 N4 ↓ Ranking	N4 N3 N1 N2 ↓ Ranking

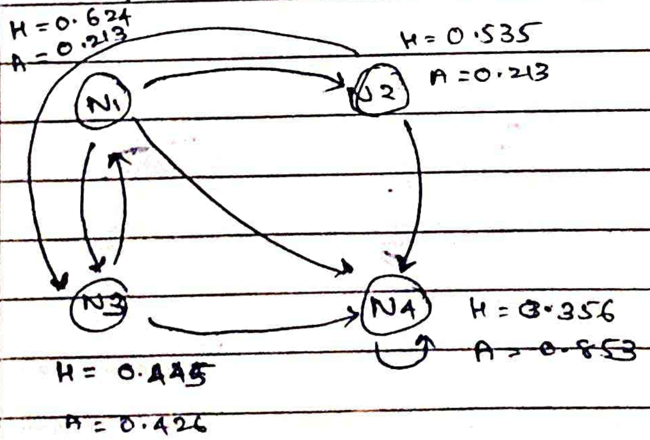
Iteration 2

$$u_1 = \begin{bmatrix} 7 \\ 6 \\ 5 \\ 4 \end{bmatrix} \qquad v_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 4 \end{bmatrix}$$

$x = 7^2 + 6^2 + 5^2 + 4^2$ $\qquad$ $y = 1^2 + 1^2 + 2^2 + 4^2$

$x = 126$ $\qquad$ $y = 22$

$$u_2 = \begin{bmatrix} 7/\sqrt{x} \\ 6/\sqrt{x} \\ 5/\sqrt{x} \\ 4/\sqrt{x} \end{bmatrix} = \begin{bmatrix} 0.624 \\ 0.535 \\ 0.445 \\ 0.356 \end{bmatrix} \qquad v_2 = \begin{bmatrix} 1/\sqrt{x} \\ 1/\sqrt{x} \\ 2/\sqrt{x} \\ 4/\sqrt{x} \end{bmatrix} = \begin{bmatrix} 0.213 \\ 0.213 \\ 0.426 \\ 0.853 \end{bmatrix}$$



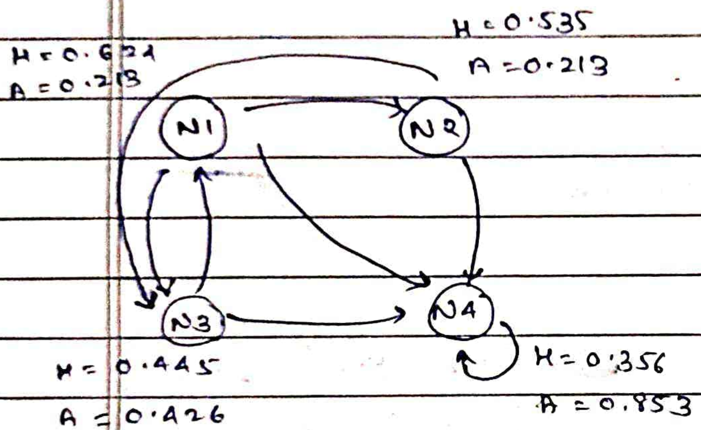
Nodes	Hub	Authority
N1	0.624	0.213
N2	0.535	0.213
N3	0.445	0.426
N4	0.356	0.853
	N1 N2 N3 N4 ↓ Ranking	N 4 N3 N1 N2 ↓ Ranking

Iteration ③ ~~[illegible]~~ =

$x = 0.624^2 + 0.535^2 + 0.445^2 + 0.356^2$ $y = 0.213^2 + 0.213^2 + 0.426^2 + 0.853$

$x = 1$ $y = 1$

$$u_3 = \begin{bmatrix} 0.624/\sqrt{x} \\ 0.535/\sqrt{x} \\ 0.445/\sqrt{x} \\ 0.356/\sqrt{x} \end{bmatrix} = \begin{bmatrix} 0.624 \\ 0.535 \\ 0.445 \\ 0.356 \end{bmatrix} \qquad v_3 = \begin{bmatrix} 0.213/\sqrt{y} \\ 0.213/\sqrt{y} \\ 0.426/\sqrt{y} \\ 0.853/\sqrt{y} \end{bmatrix} = \begin{bmatrix} 0.213 \\ 0.213 \\ 0.426 \\ 0.853 \end{bmatrix}$$



Nodes	Hub	Authority
N1	0.624	0.213
N2	0.535	0.213
N3	0.445	0.426
N4	0.356	0.853
	N1 N2 N3 N4	N4 N3 N1 N2
	↓	↓
	Ranking	Ranking