Shubham Bhate 8318 TE Comps
FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING

DWM Assignment 4

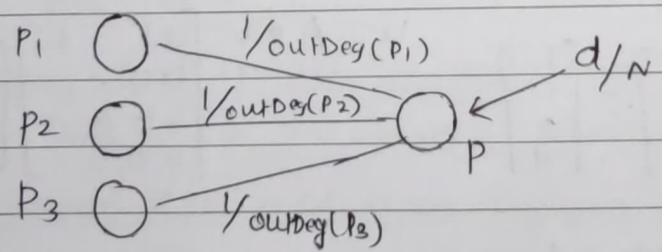
1. (i) Page rank algorithm is used by google to rank web pages in their search engine results.
(ii) PR technique is used by google to prioritize pages returned from search by looking at web structure. Importance of page is calculated based on no. of pages which point to it - Back links. Weighting is used to provide more importance
(iii) Page Rank is defined as PR(P)

$$PR(p) = c\,(PR(1)/N_1 + \cdots PR(n)/N_n)$$

PR(i): Page Rank for page i which points to target
Ni : no of links coming out of that page P
c: constant (0-1) used for normalization.

eg. Google's Page Rank
Rank of a web page depends on rank of web page pointing to it

$$PR(P) = \frac{d}{N} + (1-d)\left[\frac{PR(P_1)}{OutDeg(P_1)} + \frac{PR(P_2)}{OutDeg(P_2)} + \frac{PR(P_3)}{OutDeg(P_3)}\right]$$



set PR → [r1, r2 -- rn] these are same initial rank of pg1 to N. d = 0.15, dD ⇒ $[\frac{1}{N} \cdots \frac{1}{N})^T$
A is adjacency matrix
do $PR_{i+1} \leftarrow A^T * PR_i$
$PR_{i+1} \leftarrow (1-d) * PR_{i-1} + d * D$

$b \leftarrow ||PR_{i+1} - PR_i||$, while $b < \epsilon$ where ϵ is small no. indicating convergence threshold

return PR

2. (i) HITS (Hyperlink induced topic search) algo is a kind of rank of algo that analyzes web resources based on local link Algorithm is as follows

① Use text keyword matching to acquire root set R, which includes thousands of URL or more.

② Define S to R i.e S & R are equal

③ To each page P in R, put hyperlinks included by R into set S put pages referring to P into set S

④ S is acquired expanded neighbourhood set. The inputs for HITS algo are a set of nodes with an adjacency matrix and a value k (no of iteration). First we need to define 2 variables u: Hubs weighted vector and v: Authority weighted vector where A: adjacency matrix of all pages

$u = A * V$ and $v = A^T * u$

To calculate v we assume $u = 1$ initially and calculate v & then using this v we calculate u. Based on u, v we get the ranking for hub and authentication

For eg A is given as

	N_1	N_2	N_3	N_4
N_1	0	1	1	1
N_2	0	0	1	1
N_3	1	0	0	1
N_4	0	0	0	1

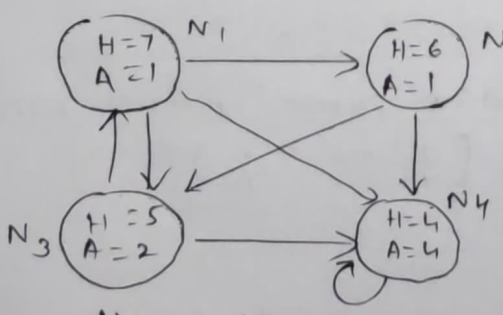
with nodes N_1, N_2, N_3, N_4, $k = 2$

Initial weight

$$u = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$V = A^T \cdot U = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 4 \end{bmatrix}$$

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



For k = 1

Nodes	Hub	Authority
N_1	7	1
N_2	6	1
N_3	5	2
N_4	4	4

Hub: N_1, N_2, N_3, N_4

Authority: $N_4, N_3, \{N_2, N_1$ TIE$\}$

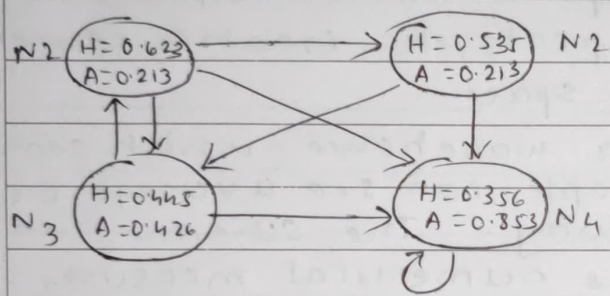
Now for k = 2, we have to use unit matrix of u, v

$|v| = \sqrt{1^2+1^2+2^2+4^2} = \sqrt{22}$

$v' = 1/\sqrt{22}, 1/\sqrt{22}, 2/\sqrt{22}, 4/\sqrt{22}$, $v' = \begin{bmatrix} 0.213 \\ 0.213 \\ 0.426 \\ 0.853 \end{bmatrix}$

similarly $|v| = \sqrt{7^2+6^2+5^2+4^2} = \sqrt{126}$

$u' = \frac{7}{\sqrt{126}}, \frac{6}{\sqrt{126}}, \frac{5}{\sqrt{126}}, \frac{4}{\sqrt{126}}$ $u' = \begin{bmatrix} 0.623 \\ 0.535 \\ 0.445 \\ 0.356 \end{bmatrix}$



For k = 2

	Hub	Authority
N1	0.623	0.213
N2	0.535	0.213
N3	0.445	0.426
N4	0.356	0.853

Ranking in same as k=1

3. (i) Multimedia database consists of audio, video, images and text media. They can be stored on object-oriented database. They are used to store complex info in a pre specified format

(ii) They store geographical info in spatial database. Stores data in form of co-ordinates, topology, lines, polygon, etc.

(iii) Time series databases : Time series databases contains stock exchange data and user logged activities. Handles arrays of nos indexed by time, date, etc.

(iv) WWW : It is a collection of documents and resources like audio, video, text, etc which are identified by URL through web browsers, linked by HTML pages and accessible via Internet Network.

*Issues in web mining:
·size of web is very huge and rapidly increasing. Complexity of webpages; web pages dont follow orders of traditional text documents. web is a dynamic information source. Data such as news, stock market, sports, etc are continuously changing. Diversity of use communities: community is rapidly expanding and have diff. backgrounds & interest. Relevance of info: It is considered that a particular person is intersted only in a small portion of the web, while the rest is not relevant to the user.

* Issues in spatial mining:
Accurately representing true shape and size. Representing non-continuous data eg. road lines, peaks, etc. Creating aesthetically pleasing maps. conserving disk space.

4.] we can develop spatial data warehouse which stores highway traffic info so that people can see average or peak time traffic flow by the highway. The schema can be a star with flow and count as numerical measures. The region map measure is spatial and represents collection of spatial pointers to corresponding regions. So based on this data warehouses, as a city planner of Mumbai, different kinds of information can be mined. For eg. he can notice where max traffic is generated & broaden the roads or divert the vehicles. We can use mining to learn live traffic updates and no. of accidents at a location. Such info is useful for planning infrastructure.

5.] It is unrealistic to treat video, as a long sequence of individual stills and analyze each picture. This would result in too many pictures. In order to capture outliers from video data, it is better to treat each video clip first as a collection of actions and events in time, segment them into video shots. A shot is a group of frames where video content from one frame to adjacent ones does not change abruptly. Each frame can be analyzed using image feature extraction and other methods.. A process can then sample the data quickly and analyze images in real time. The process compares features of key frame with current static images to see if there is a major difference. If so then alert an alarm that there is some unusual event occurring.