Advanced Data Analysis

Link Analysis, Frequent item set mining and Hierarchical Clustering

Link Analysis

- A collection of techniques that can be applied to data having relationships among themselves

- Centrality
  - Degree
  - Closeness
  - Betweeness

- Prestige
  - Page Rank Algorithm

  References:

- https://www.youtube.com/watch?v=7tRxCpHhDcw&t=1373s

Page Rank

- PageRank relies on the democratic nature of the web by using its vast link structures as an indicator of an individual page’s value or quality.
- It interprets a hyperlink from page x to page y as a vote, by page x, for page y.
- However, page rank looks at more than sheer number of votes; it analyzes the page the casts the vote.
  - Votes casted by important pages weight more heavily and help to make other pages more important.
  - This is exactly the idea of rank prestige in social network.

Page Rank

A hyperlink from a page to another page is an implicit conveyance of authority to the target page.

- The more in-links that a page *i* receives, the more prestige the page *i* has.

Pages that point to page *i* also have their own prestige scores.

- A page of a higher prestige pointing to *i* is more important than a page of a lower prestige pointing to *i*.
- In other words, a page is important if it is pointed to by other important pages.

Page Rank Algorithm

According to rank prestige, the importance of page i (i's PageRank score) is the sum of the PageRank scores of all pages that point to i.

Since a page may point to many other pages, its prestige score should be shared.

The Web as a directed graph $G = (V, E)$. Let the total number of pages be n. The PageRank score of the page i (denoted by $P(i)$) is defined by:

$$P(i) = \sum_{(j,i) \in E} \frac{P(j)}{O_j}$$

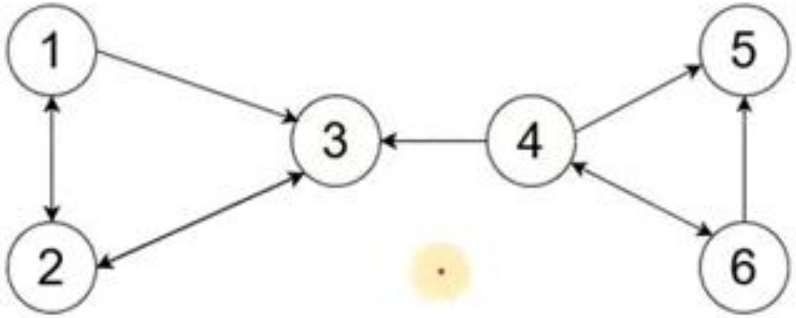
$$P_{i+1} = A^T P_i$$

To introduce these conditions and the enhanced equation, let us derive the same Equation based on the Markov chain.

- In the Markov chain, each Web page or node in the Web graph is regarded as a state.
- A hyperlink is a transition, which leads from one state to another state with a probability.

Page Rank

$$A = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$



Page Rank

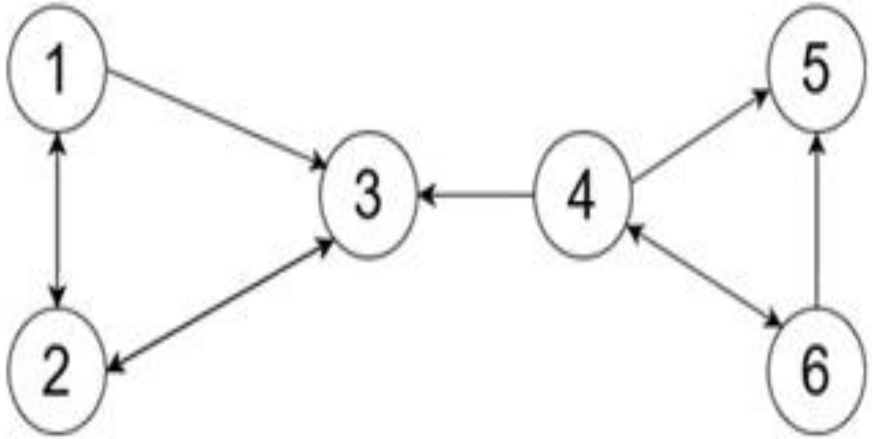
FIX THE PROBLEM: TWO POSSIBLE WAYS

1. **Remove those pages with no out-links during the PageRank computation as these pages do not affect the ranking of any other page directly.**
2. **Add a complete set of outgoing links from each such page *i* to all the pages on the Web.**

Let us use the second way

$$\overline{A} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Page Rank



$$\overline{A} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

$$P_{i+1} = A^T P_i$$

Page Rank

$$A^T = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/6 & 0 \\ 1/2 & 0 & 1 & 0 & 1/6 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 0 \end{pmatrix}$$

$$P_o = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}$$

Page Rank

$$P_1 = A^T P_o$$

$$P_1 = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/6 & 0 \\ 1/2 & 0 & 1 & 0 & 1/6 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 1.833 \\ 4.333 \\ 3.666 \\ 3.833 \\ 5.166 \\ 2.166 \end{pmatrix}$$

$$P_2 = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/6 & 0 \\ 1/2 & 0 & 1 & 0 & 1/6 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 0 \end{pmatrix} \begin{pmatrix} 1.833 \\ 4.333 \\ 3.666 \\ 3.833 \\ 5.166 \\ 2.166 \end{pmatrix} = \begin{pmatrix} 3.021 \\ 5.433 \\ 5.212 \\ 1.937 \\ 3.213 \\ 2.132 \end{pmatrix}$$

Page Rank

$$P_3 = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/6 & 0 \\ 1/2 & 0 & 1 & 0 & 1/6 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 0 \end{pmatrix} \begin{pmatrix} 3.021 \\ 5.433 \\ 5.212 \\ 1.937 \\ 3.213 \\ 2.132 \end{pmatrix} = \begin{pmatrix} 3.250 \\ 7.256 \\ 5.406 \\ 1.599 \\ 2.244 \\ 1.178 \end{pmatrix}$$

$$P_9 = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/6 & 0 \\ 1/2 & 0 & 1 & 0 & 1/6 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 0 \end{pmatrix} \begin{pmatrix} 4.518 \\ 8.898 \\ 6.784 \\ 0.210 \\ 0.315 \\ 0.182 \end{pmatrix} = \begin{pmatrix} 4.501 \\ 9.096 \\ 6.830 \\ 0.143 \\ 0.213 \\ 0.122 \end{pmatrix}$$

Rank: 2, 3, 1, 5, 4, 6

#Experiment 1

P0 = [1, 2, 3, 4, 5, 6]

P9 = [4.501, 9.096 6.830 0.143 0.213 0.122]

Ranking: [2, 3, 1, 5, 4, 6]

#Experiment 2

P0 = [4, 3, 6, 1, 5, 2]

P9 = [4.559 9.247 6.914 0.067 0.100 0.057]

Ranking: [2, 3, 1, 5, 4, 6]

#Experiment 3

P0 = [100, 100, 100, 100, 100, 100]

P9 = [130.772 261.197 196.792 2.810 4.188 2.405]

Page Rank

#Experiment 4

P0 = [0.166, 0.166, 0.166, 0.166, 0.166, 0.166]

P1 = [0.110	0.276	0.248	0.110	0.165	0.082]
P2 = [0.165	0.331	0.257	0.068	0.105	0.064]
P3 = [0.183	0.358	0.289	0.049	0.072	0.040]
P4 = [0.191	0.392	0.299	0.032	0.048	0.028]
P5 = [0.204	0.403	0.310	0.022	0.033	0.018]
P6 = [0.207	0.418	0.316	0.014	0.022	0.012]
P7 = [0.213	0.424	0.321	0.010	0.015	0.008]
P8 = [0.214	0.430	0.324	0.006	0.010	0.005]
P9 = [0.217	0.433	0.326	0.004	0.006	0.003]

Frequent Item Sets Analysis

- Apriori Algorithm
- FP Growth Algorithm

Hierarchical Clustering

- Page 245, Book