Week 4:
Web Mining: Google PageRank

CMPS 620 : Machine Learning

Introduction

- PageRank is the first algorithm used by Google in sorting search results.
- It was developed at Stanford University by two doctoral computer science students, Larry Page and Sergey Brin, in 1996 as part of a research project about a new kind of search engine.
- The name "PageRank" plays off of the name of developer Larry Page, as well as the concept of a web page.
- There are other search algorithms available now, such as Google Panda.
  - Currently, Google uses an automated web spider called **Googlebot** to actually count links and gather other information on web pages.

Main Idea of PageRank

- Given a collection of n web pages, the goal is to rank them in terms of importance.
- Roughly speaking, PageRank represents the probability that a person randomly clicking on links will arrive at any particular page.
- The mathematics involved in PageRank includes decomposition of positive or nonnegative matrices and homogeneous Markov Chain in stochastic process.

Linked Web Pages

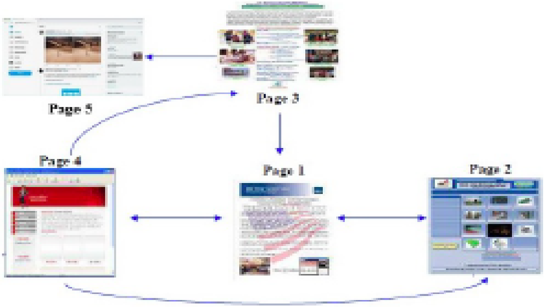


Figure: Illustration of PageRank

- Page 5 is called the 'dangling node' as it has no outlink

Graphical Representation

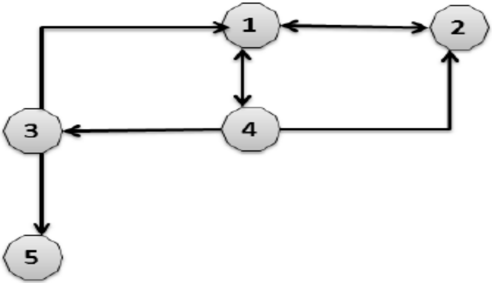


Figure: Illustration of PageRank

- Page 5 is called the ‘dangling node’ as it has no outlink

Factors under Consider

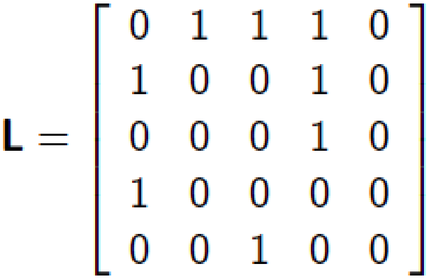
- There are two major factors to consider when ranking pages.
  1. First of all, a webpage is important if many other webpages point to it.
  2. However the linking webpages that point to a given page are not treated equally:
     - ★ the algorithm takes into account both the importance (PageRank) of the linking pages and the number of outgoing links that they have.
- Linking pages with higher PageRank are given more weight, while pages with more outgoing links are given less weight.
- These ideas lead to a recursive definition for PageRank.

The Link Matrix

- Let $L_{ij} = 1$ if page j points to page i ; and zero otherwise.
- Let $c_j = \sum_{i=1}^{n} L_{ij}$ (the column sums) equal the number of pages pointed to by page j (number of 'outlinks' or 'forward links').
- Similarly, the row sum $r_j = \sum_{j=1}^{n} L_{ij}$ represents the number of 'back links'.

The Link Matrix

- For example, the link matrix L with Figure 1 is given below:



$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

and $c = (2, 1, 2, 3, 0)^T$.

Defining PageRank

Definition:
The Google PageRank p_i is defined recursively by,

$$\begin{aligned} p_i =& \sum_{j=1}^{n} L_{ij} p_j \text{ concerning factor 1} \\ =& \sum_{j=1}^{n} \frac{L_{ij}}{c_j} p_j \text{ concerning factor 1 and 2} \\ =& (1-d) + d \cdot \sum_{j=1}^{n} \frac{L_{ij}}{c_j} p_j \end{aligned}$$

with some $0 < d < 1$

The Damping Factor

- In the last step, a damping factor d is introduced.
- Mathematically, this modification does not alter the nature of the transition matrix A (definition will be given shortly) as being a transition matrix in homogeneous Markov Chain.
- Page et al. (1998) considered PageRank as a model of user behavior, where a random web surfer clicks on links at random, without regard to content.
    - The surfer does a random walk on the web, choosing among available outgoing links at random.
- The factor $1 - d$ is the probability that he does not click on a link, but jumps instead to a random webpage.
    - In PageRank, d is arbitrarily set as .85.

The Damping Factor

In matrix form, let $\mathbf{e} = (\mathbf{1})_\mathbf{n}$, $\mathbf{J} = \mathbf{e}\mathbf{e}^\mathbf{T}$, $\mathbf{c} = (\mathbf{c_j})$, and $\mathbf{D_c} = \mathbf{diag}(\mathbf{c})$. Then it follows that

$$\mathbf{p} = (1-d)\mathbf{e} + d\mathbf{L}\mathbf{D}_\mathbf{c}^{-\mathbf{1}}\mathbf{p} \tag{1}$$

A further normalization is introduced so that the average PageRank is 1. In other words, $\mathbf{e}^\mathbf{T}\mathbf{p} = n$. Hence, (1) becomes

$$\mathbf{p} = \{(1-d)\mathbf{J}/n + d\mathbf{L}\mathbf{D}_\mathbf{c}^{-\mathbf{1}}\}\mathbf{p} = \mathbf{A}\mathbf{p} \tag{2}$$

Solving for PageRank **p** - Approach I

The first approach for solving p is algebraic:

- Matrix **A** is nonnegative with $a_{ij} \geq 0$:
- Perron-Frobenius theorem provides the foundation for decomposing a nonnegative matrix.
- Since $1 \cdot \mathbf{p} = \mathbf{Ap}$, it can be shown that **A** has the first eigenvalue $\lambda_1 = 1$ with multiplicity 1 (meaning all other $\lambda < 1$).
- Thus **p** is the so-called Perron (right) vector of **A** and can be obtained via the decomposition of **A**.

Solving for PageRank $\mathbf{p}$ - Approach II

The second approach gains insight from stochastic process:

- Matrix $\mathbf{A}$ (which has nonnegative entries with each column summing to one) is a stochastic matrix for a homogeneous Markov Chain (HMC).
- The solution $\mathbf{p}$ can be found via the power method by iterating between $\mathbf{p}_k = \mathbf{A}\mathbf{p}_{k-1}$ and normalization $\mathbf{p}_k = n\frac{\mathbf{p}_k}{\mathbf{e}^\mathsf{T}\mathbf{p}_k}$.