

Introduction to Data Mining

Lecture #9: Link Analysis

U Kang Seoul National University

U Kang





➡ □ Overview

PageRank: Flow Formulation



Graph Data: Social Networks



Facebook social graph

4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011] **U** Kang



Graph Data: Media Networks



Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]



Graph Data: Information Nets



Graph Data: Communication Nets





Graph Data: Classic Example



Seven Bridges of Königsberg

[Euler, 1735]

Return to the starting point by traveling each link of the graph once and only once.

U Kang





Web as a Graph

Web as a directed graph:

Nodes: Webpages

Edges: Hyperlinks





Web as a Graph

Web as a directed graph:

- Nodes: Webpages
- Edges: Hyperlinks



9



Web as a Directed Graph





Broad Question

How to organize the Web?

- First try: Human curated
 Web directories
 - Yahoo, DMOZ, LookSmart
- Second try: Web Search
 - Information Retrieval investigates:
 Find relevant docs in a small and trusted set
 - Newspaper articles, Patents, etc.
 - <u>But:</u> Web is huge, full of untrusted documents, random things, web spam, etc.

Now Open (shoo) 2nd Shop	Play (100 Web Lauch	
	Search Options	
Arts Humatian, Photography, Architecture,	News (Strat) Voil (Strat), Duby, Const Event,	
Business and Economy(Erst) Directory, Investment, Charliele, Tame,	Recreation Spots [Brad], Gamer, Torol, Aster,	
Computers and Internet(Bowl) herest, VVV, Solvers, Millinelle,	Reference Liberter, Distances, Phone Humber,	
Education Universities, E-12, Courses,	• Regional Countier, Regione, U.S. Rober,	
Entertainment (20val) TV, Movier, Music, Mappiner,	• Science CR, Einley, Advance, Espinetic,	
Government Felicie (Steaf), Aparter, Lev, Miltery,	Social Science Anthropology, Bening, Desamine,	
· Health	· Society and Culture	



Web Search: 2 Challenges

- 2 challenges of web search:
- (1) Web contains many sources of information Who to "trust"?

Idea: Trustworthy pages may point to each other!

- (2) What is the "best" answer to the query "newspaper"?
 - No single right answer
 - Idea: Pages that actually know about newspapers might all be pointing to many newspapers



Ranking Nodes on the Graph

All web pages are not equally "important"

www.joe-schmoe.com vs. www.snu.ac.kr

 There is large diversity in the web-graph node connectivity.
 Let's rank the pages by the link structure!





Link Analysis Algorithms

- We will cover the following Link Analysis approaches for computing importances of nodes in a graph:
 - Page Rank
 - Topic-Specific (Personalized) Page Rank
 - Web Spam Detection Algorithms







➡ □ PageRank: Flow Formulation



Links as Votes

Idea: Links as votes

A page is more important if it has more links

In-coming links? Out-going links?

Think of in-links as votes:

- www.snu.ac.kr has 100,000 in-links
- www.joe-schmoe.com has 1 in-link

Are all in-links equal?

- Links from important pages count more
- Recursive question!



Example: PageRank Scores





Simple Recursive Formulation

- Each link's vote is proportional to the importance of its source page
- If page j with importance r_j has n out-links, each link gets r_j / n votes
- Page j's own importance is the sum of the votes on its in-links

$$r_j = r_i/3 + r_k/4$$



PageRank: The "Flow" Model

- A "vote" from an important page is worth more
- A page is important if it is pointed to by other important pages
- Define a "rank" r_j for page j

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

d_i ... out-degree of node *i* i -> j : all i that point to j U Kang



"Flow" equations: $r_y = r_y/2 + r_a/2$ $r_a = r_y/2 + r_m$ $r_m = r_a/2$



Solving the Flow Equations

3 equations, 3 unknowns, no constants

- No unique solution
- All solutions equivalent modulo the scale factor
 $r_m = r_a/2$
 - I.e., Multiplying c to given a solution r_y, r_a, r_m will give you another solution

Additional constraint forces uniqueness:

$$r_y + r_a + r_m = 1$$

• Solution:
$$r_y = \frac{2}{5}$$
, $r_a = \frac{2}{5}$, $r_m = \frac{1}{5}$

- Gaussian elimination method works for small examples, but we need a better method for large web-size graphs
- We need a new formulation!

Flow equations:

 $r_v = r_v / 2 + r_a / 2$

 $r_a = r_v/2 + r_m$

PageRank: Matrix Formulation

U Kang

Stochastic adjacency matrix M

• Let page i has d_i out-links

If
$$i \to j$$
, then $M_{ji} = \frac{1}{d}$ else M_{ji}

M is a column stochastic matrix
 Columns sum to 1

NOTE: A matrix **M** is called `column stochastic' if the sum of each column is 1







PageRank: Matrix Formulation

Rank vector r: vector with an entry per page
r_i is the importance score of page i
\sum_i r_i = 1
The flow equations $r_j = \sum_{i o j} \frac{r_i}{d_i}$ can be written $r = M \cdot r$

Why?





Eigenvector Formulation

- The flow equations can be written
 - $r = M \cdot r$

NOTE1: \boldsymbol{x} is an eigenvector with the corresponding eigenvalue $\boldsymbol{\lambda}$ if:

- Ax = λx
 So the rank vector r is an eigenvector of the web matrix M, with the corresponding eigenvalue 1
- Fact: The largest eigenvalue of a column stochastic matrix is 1
- We can now efficiently solve for r! The method is called Power iteration



Example: Flow Equations & M



	У	a	m
У	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

 $r = M \cdot r$

$$r_{y} = r_{y}/2 + r_{a}/2$$
$$r_{a} = r_{y}/2 + r_{m}$$
$$r_{m} = r_{a}/2$$

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$



Power Iteration Method

- Given a web graph with n nodes, where the nodes are pages and edges are hyperlinks
- Power iteration: a simple iterative scheme
 - □ Suppose there are *N* web pages
 - □ Initialize: $\mathbf{r}^{(0)} = [1/N, ..., 1/N]^{T}$
 - **u** Iterate: $\mathbf{r}^{(t+1)} = \mathbf{M} \cdot \mathbf{r}^{(t)}$
 - Stop when $|\mathbf{r}^{(t+1)} \mathbf{r}^{(t)}|_1 < \varepsilon$ $|\mathbf{x}|_1 = \sum_{1 \le i \le N} |\mathbf{x}_i|$ (called the L₁ norm) Can use any other vector norm, e.g., Euclidean



d_i out-degree of node i



Power Iteration Method

Power iteration:

A method for finding dominant eigenvector (the vector corresponding to the largest eigenvalue)

$$\neg r^{(1)} = M \cdot r^{(0)}$$

$$r^{(2)} = M \cdot r^{(1)} = M(Mr^{(1)}) = M^2 \cdot r^{(0)}$$
$$r^{(3)} = M \cdot r^{(2)} = M(M^2r^{(0)}) = M^3 \cdot r^{(0)}$$

Sequence $M \cdot r^{(0)}, M^2 \cdot r^{(0)}, ... M^k \cdot r^{(0)}, ...$ approaches the dominant eigenvector of M

 Dominant eigenvector = the one corresponding to the largest eigenvalue



PageRank: How to solve?

a

Power Iteration:

• Set $r_j = 1/N$ • 1: $r'_j = \sum_{i \to j} \frac{r_i}{d_i}$

2:
$$r = r'$$

Goto 1

Example:





 $r_{y} = r_{y}/2 + r_{a}/2$ $r_{a} = r_{y}/2 + r_{m}$ $r_{m} = r_{a}/2$

lKang

Random Walk Interpretation

- Imagine a random web surfer:
 - At any time t, surfer is on some page i
 - At time t + 1, the surfer follows an out-link from i uniformly at random
 - Ends up on some page *j* linked from *i*
 - Process repeats indefinitely
- Let:
 - p(t) ... vector whose i^{th} coordinate is the prob. that the surfer is at page i at time t
 - So, p(t) is a probability distribution over pages







The Stationary Distribution

• Where is the surfer at time *t*+*1*?

Follows a link uniformly at random

 $p(t+1) = M \cdot p(t)$



• Suppose the random walk reaches a state $p(t + 1) = M \cdot p(t) = p(t)$ then p(t) is called stationary distribution of a ran

then p(t) is called stationary distribution of a random walk

Our original rank vector r satisfies r = M · r
 So, r is a stationary distribution for the random walk



Existence and Uniqueness

A central result from the theory of random walks (a.k.a. Markov processes):

For graphs that satisfy **certain conditions**, the **stationary distribution is unique** and eventually will be reached no matter what the initial probability distribution at time **t** = **0**

Certain conditions: a walk starting from a random page can reach any other page



Questions?