Analysis of Google's PageRank

Ilse Ipsen

North Carolina State University

Joint work with Rebecca M. Wills

PageRank

An objective measure of the citation importance of a web page [Brin & Page 1998]

- Assigns a rank to every web page
- Influences the order in which Google displays search results
- Based on link structure of the web graph
- Topic independent

Overview

- Google Matrix
- Stability of PageRank
- Eigenvalue Problem: Power Method
- Linear System: Jacobi Method
- Dangling Nodes

Simple Web Model

Construct matrix S

- Page i has $d \geq 1$ outgoing links:
  If page i has link to page j then $s_{ij} = 1/d$
  else $s_{ij} = 0$
- Page i has 0 outgoing links: $s_{ij} = 1/n$
  (dangling node)

s_{ij}: probability that surfer moves
from page i to page j

Google Matrix

Convex combination

$$G = \alpha S + (1 - \alpha) 1 v^T$$

Stochastic matrix S
Damping factor $0 < \alpha < 1$, e.g. $\alpha = .85$
Personalization vector $v \geq 0 \quad \|v\|_1 = 1$

TrustRank (to combat web spam)

$v_i = 0$ if page i is spam page

[Gyöngyi, Garcia-Molina, Pedersen 2004]

Properties of G

$$G = \alpha S + (1 - \alpha) 1 v^T$$

- Stochastic, reducible
- Eigenvalues: $1 > \alpha\lambda_2(S) \geq \alpha\lambda_3(S) \geq \ldots$
  [Elden 2003]
- Unique left eigenvector:

$$\pi^T G = \pi^T \qquad \pi \geq 0 \qquad \|\pi\|_1 = 1$$

ith entry of π: PageRank of page i

PageRank $\doteq$ largest left eigenvector of G

Stability of PageRank

How sensitive is PageRank π to

- Round off errors
- Changes in damping factor α
- Changes in personalization vector v
- Addition/deletion of links

Perturbation Theory

For Markov chains

Schweizer 1968, Meyer 1980
Haviv & van Heyden 1984
Funderlic & Meyer 1986
Golub & Meyer 1986
Seneta 1988, 1991
Ipsen & Meyer 1994
Kirkland, Neumann & Shader 1998
Cho & Meyer 2000, 2001
Kirkland 2003, 2004

Perturbation Theory

For Google matrix

Chien, Dwork, Kumar & Sivakumar 2001
Ng, Zheng & Jordan 2001
Bianchini, Gori & Scarselli 2003
Boldi, Santini & Vigna 2004
Langville & Meyer 2004
Golub & Greif 2004
Kirkland 2005
Chien, Dwork, Kumar, Simon & Sivakumar 2005

Changes in the Matrix S

Exact:

$$\pi^T G = \pi^T \qquad G = \alpha S + (1-\alpha) 1 v^T$$

Perturbed:

$$\tilde{\pi}^T \tilde{G} = \tilde{\pi}^T \qquad \tilde{G} = \alpha(S + E) + (1-\alpha) 1 v^T$$

Error:

$$\tilde{\pi}^T - \pi^T = \alpha \tilde{\pi}^T E (I - \alpha S)^{-1}$$

$$\|\tilde{\pi} - \pi\|_1 \leq \frac{\alpha}{1-\alpha} \|E\|_\infty$$

Changes in Damping Factor α

Exact:

$$\pi^T G = \pi^T \qquad G = \alpha S + (1-\alpha) 1 v^T$$

Perturbed:

$$\tilde{\pi}^T \tilde{G} = \tilde{\pi}^T \qquad \tilde{G} = (\alpha+\mu) S + (1-(\alpha+\mu))\, 1 v^T$$

Error:

$$\|\tilde{\pi} - \pi\|_1 \leq \frac{2}{1-\alpha} |\mu|$$

[Langville & Meyer 2004]

Changes in Vector v

Exact:

$$\pi^T G = \pi^T \qquad G = \alpha S + (1-\alpha)1v^T$$

Perturbed:

$$\tilde{\pi}^T \tilde{G} = \tilde{\pi}^T \qquad \tilde{G} = \alpha S + (1-\alpha)1(v+f)^T$$

Error:

$$\|\tilde{\pi} - \pi\|_1 \leq \|f\|_1$$

Sensitivity of PageRank π

$$\pi^T G = \pi^T \qquad G = \alpha S + (1-\alpha) 1 v^T$$

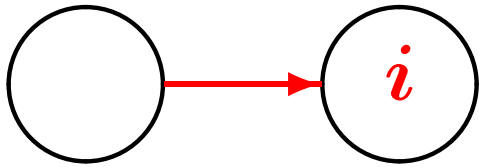
Changes in

- S: condition number $\alpha/(1-\alpha)$
- α: condition number $2/(1-\alpha)$
- v: condition number 1

$\alpha = .85$: condition numbers ≤ 14
$\alpha = .99$: condition numbers ≤ 200

PageRank insensitive to perturbations

Adding an In-Link



$$\tilde{\pi}_i > \pi_i$$

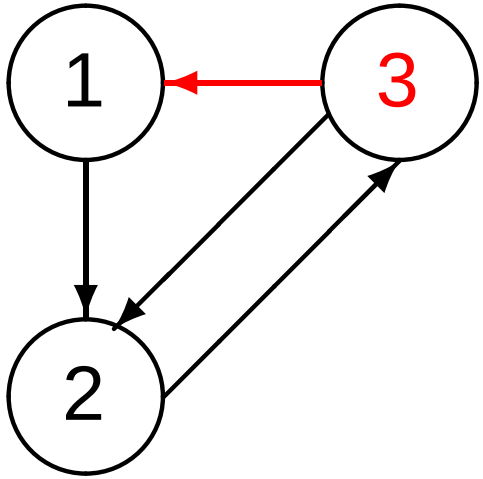
Adding an in-link increases PageRank (monotonicity)

Removing an in-link decreases PageRank

[Chien, Dwork, Kumar & Sivakumar 2001]

[Chien, Dwork, Kumar, Simon & Sivakumar 2005]

Adding an Out-Link



$$\tilde{\pi}_3 = \frac{1+\alpha+\alpha^2}{3(1+\alpha+\alpha^2/2)} < \pi_3 = \frac{1+\alpha+\alpha^2}{3(1+\alpha)}$$

Adding an out-link may decrease PageRank

Justification for TrustRank

Adjust personalization vector to change PageRank

Increase v for page i: $v_i := v_i + \phi$
Decrease v for page j: $v_j := v_j - \phi$

PageRank of page i increases: $\tilde{\pi}_i > \pi_i$
PageRank of page j decreases: $\tilde{\pi}_j < \pi_j$

Total change in PageRank $\|\tilde{\pi} - \pi\|_1 \leq 2\phi$

PageRank Computation

- Power method
  Page, Brin, Motwani & Winograd 1999
- Acceleration of power method
  Kamvar, Haveliwala, Manning & Golub 2003
  Haveliwala, Kamvar, Klein, Manning & Golub 2003 Brezinski & Redivo-Zaglia 2004
  Brezinski, Redivo-Zaglia & Serra-Capizzano 2005
- Aggregation/Disaggregation
  Langville & Meyer 2002, 2003, 2004
  Ipsen & Kirkland 2004

PageRank Computation

- Methods that adapt to web graph
  Broder, Lempel, Maghoul & Pedersen 2004
  Kamvar, Haveliwala & Golub 2004
  Haveliwala, Kamvar, Manning & Golub 2003
  Lee, Golub & Zenios 2003
  Lu, Zhang, Xi, Chen, Liu, Lyu & Ma 2004
- Krylov methods
  Golub & Greif 2004

Power Method

Eigenvector problem: $\pi^T G = \pi^T$

Power method:

Pick $x^{(0)} > 0 \quad \|x^{(0)}\|_1 = 1$

Repeat $[x^{(k+1)}]^T = [x^{(k)}]^T G$

until $\|x^{(k+1)} - x^{(k)}\| \leq \tau$

$[x^{(k+1)}]^T - [x^{(k)}]^T = [x^{(k)}]^T G - [x^{(k)}]^T$ residual

Error in Power Method

$$\pi^T G = \pi^T \qquad G = \alpha S + (1-\alpha) 1 v^T$$

Error in iteration k

- Forward: $e_k \equiv x^{(k)} - \pi$

$$\|e_k\|_{1,\infty} \leq \alpha^k \, \|e_0\|_{1,\infty}$$

[Bianchini, Gori & Scarselli 2003]

- Residual: $r_k^T = [x^{(k)}]^T G - [x^{(k)}]^T$

$$\|r_k\|_{1,\infty} \leq \alpha^k \, \|r_0\|_{1,\infty}$$

Termination

Residual norm $\|r_k\|_\infty \leq 2\alpha^k$

Stop when $\|r_k\|_\infty \leq 10^{-8}$

For $\alpha = .85$: $k \geq 119$

n	2293	2947	3468	5757	281903	683446
k	75	76	83	79	69	65

Bound can be pessimistic

PageRank from Linear System

Eigenvector problem:

$$\pi^T \underbrace{(\alpha S + (1-\alpha) 1 v^T)}_{G} = \pi^T \qquad \pi \geq 0 \quad \|\pi\|_1 = 1$$

Linear system:

$$\pi^T (I - \alpha S) = (1-\alpha) v^T$$

$I - \alpha S$ nonsingular M-matrix

[Arasu, Novak, Tomkins & Tomlin 2002]

[Bianchini, Gori & Scarselli 2003]

Jacobi Method

Assume no page has a link to itself

$$\pi^T(I - \alpha S) = (1 - \alpha)v^T \qquad I - \alpha S = D - O$$

$$[x^{(k+1)}]^T = [x^{(k)}]^T O D^{-1} + (1 - \alpha)v^T D^{-1}$$

- $I - \alpha S$ is M-matrix
- Jacobi converges
- No dangling nodes: $D = I \quad O = \alpha S$
  Jacobi method = power method

Dangling Nodes

$S = H + dw^T$ is dense

What to do about dangling nodes?

- Remove [Brin, Page, Motwani & Winograd 1998]
  No PageRank for dangling nodes
  Biased PageRank for other nodes
- Lump into single state [Lee, Golub & Zenios 2003]
  As above
- Remove dw^T [Langville & Meyer 2004]
  [Arasu, Novak, Tomkins & Tomlin 2002]
  H is not stochastic
  What is being computed?

Use v for Dangling Nodes

$$\pi^T(I - \alpha S) = (1-\alpha)v^T \qquad S = H + dw^T$$

Choose $w = v$

$$\pi^T(I - \alpha H) = \underbrace{(1 - \alpha + \alpha\pi^T d)\, v^T}_{\text{multiple of } v^T}$$

Solve $\delta^T(I - \alpha H) = \text{multiple of } v^T$

Then δ is multiple of π

[Gleich, Zhukov & Berkhin 2005]

Iteration Counts for $w = v$

n	Power	Jacobi
2293	75	74
2947	76	74
3468	83	82
5757	79	78
281903	69	69
683446	65	65

$\|r_k\|_\infty \leq 10^{-8}$, $\alpha = .85$

Jacobi: same # iterations as power method

Extension to Arbitrary w

$$\pi^T(I - \alpha S) = (1-\alpha)v^T \qquad S = H + dw^T$$

Rank-one update: $I - \alpha S = (I - \alpha H) - \alpha d w^T$

1. Solve $\delta^T(I - \alpha H) = (1-\alpha)v^T$
2. Solve $\omega^T(I - \alpha H) = w^T$
3. Update $\pi^T = \delta^T + \frac{\alpha\, \delta^T d}{1 - \alpha\, \omega^T d}\, \omega^T$

This requires only two sparse solves

Solve with $I - \alpha H$

After similarity permutation:

$$H = \begin{pmatrix} H_1 & H_2 \\ 0 & 0 \end{pmatrix}$$

$$(x_1^T \quad x_2^T) \begin{pmatrix} I - \alpha H_1 & -\alpha H_2 \\ 0 & I \end{pmatrix} = (b_1^T \quad b_2^T)$$

1. Sparse solve $x_1^T (I - \alpha H_1) = b_1$
2. Set $x_2^T = \alpha\, x_1^T H_2 + b_2^T$

PageRank via Linear System

Rank one update: $S = \begin{pmatrix} H_1 & H_2 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w^T$

- General dangling node vector $w \geq 0$, $\|w\|_1 = 1$
  Traditional: $w = \frac{1}{n}1, \quad w = v$
- Cost:
  Two sparse solves with $I - \alpha H_1$ via Jacobi
  Two matrix vector multiplications with H_2
  Inner products and vector additions
- More dangling nodes $\Rightarrow$ cheaper

Summary

Google Matrix $G = \alpha\, S + (1 - \alpha)\, 1 v^T$

- PageRank = left eigenvector of G
- PageRank insensitive to perturbations in G
- Adding in-links increases PageRank
- Adding out-links may decrease PageRank
- Justification for TrustRank

Summary, ctd

- Power method
- Computes PageRank as eigenvector
- Forward, backward errors $\leq 2\,\alpha^k$
- Jacobi Method
- Computes PageRank from linear system
- No dangling nodes: Jacobi = power method
- Rank one update for dangling nodes
- More dangling nodes $\Rightarrow$ cheaper