Analysis and Computation of Google's PageRank

Ilse Ipsen

North Carolina State University

Joint work with Rebecca M. Wills

Overview

Goal:
Compute (citation) importance of a web page

- Simple Web Model
- 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

Matrix S

S is stochastic: $0 \le s_{ij} \le 1 \quad S\mathbf{1} = \mathbf{1}$

Left eigenvector: $\omega^T S = \omega^T \quad \omega \ge 0 \quad \|\omega\|_1 = 1$

Ranking: ω_i is probability that surfer visits page i

But:

- S does not model surfing behaviour properly
- Rank sinks, and pages with zero rank
- Several eigenvalues with magnitude 1
  $\Rightarrow$ power method does not converge

Remedy: Change the matrix

Google Matrix

Convex combination

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

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

Properties of G:

- stochastic $\Rightarrow$ G has eigenvalue 1
- primitive $\Rightarrow$ spectral radius 1 unique

Page Rank

Unique left eigenvector:

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

Power method converges to π

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
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

Changes in the Matrix S

Exact:

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

Perturbed:

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

Error:

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

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

[Kirkland 2005]

Changes in Damping Factor α

Exact:

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

Perturbed:

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

Error:

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

[Langville & Meyer 2004]

Changes in Vector v

Exact:

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

Perturbed:

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

Error:

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

Sensitivity of PageRank π

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

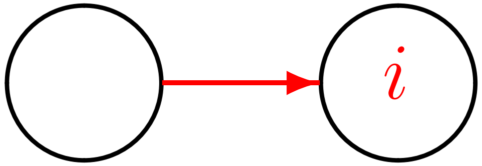
Changes in

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

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

PageRank insensitive to perturbations

Adding an In-Link



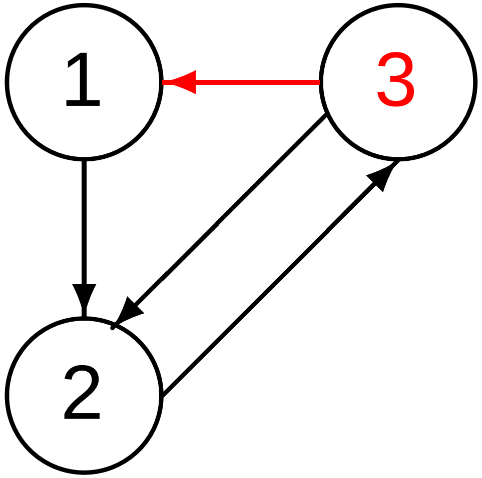
$$\tilde{\pi}_i \geq \pi_i$$

Adding an in-link can only increase PageRank

Removing an in-link can only decrease PageRank

[Chien, Dwork, Kumar, Sivakumar 2001]

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

PageRank Computation

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

Compute π by power method:

Pick $x_0 > 0$, $\|x_0\|_1 = 1$

For $k = 0, 1, 2 \ldots \qquad x_{k+1}^T = x_k^T G$

Error in iteration k:

$$\|x_k - \pi\|_1 \leq \alpha^k \, \|x_0 - \pi\|_1$$

[Bianchini, Gori & Scarselli 2003]

Why is Power Method Cheap?

Google matrix $G = \alpha S + (1-\alpha)\mathbf{1}v^T$

$$S = H + \underbrace{dw^T}_{\text{dangling nodes}} \qquad w \geq 0 \quad \|w\|_1 = 1$$

Matrix H: models webgraph
substochastic
dimension: several billion
very sparse

Matrix Vector Multiplication

Vector $x > 0 \quad \|x\|_1 = 1$

$$\begin{aligned} x^T G &= x^T \left[\alpha(H + dw^T) + (1-\alpha)\mathbf{1}v^T\right] \\ &= \alpha\, x^T H + \underbrace{\alpha\, x^T d}_{\text{scalar}}\, w^T + (1-\alpha)v^T \end{aligned}$$

Cost: # non-zeros in H

Stopping Criterion

Residual

$$x_k^T G - x_k = x_{k+1}^T - x_k^T$$

Currently: stop when

$$\|x_{k+1} - x_k\|_1 \leq \tau$$

where $\tau \approx 10^{-4}, 10^{-6}, 10^{-8}$

Better: stop when

$$\|x_{k+1} - x_k\|_1 \leq n\tau$$

where n is dimension of G

Comparison

n	its (τ)	its ($n\tau$)	Agrees	Disagrees
2293	86	39	2257	36
2947	85	37	2977	20
3468	90	41	3462	6
5757	90	39	5735	22

Its ($n\tau$): # iterations with bound $n\tau$

Agrees: # pages with same ranking for both bounds

$\tau = 10^{-8}$

Stopping Criterion

Old: $$\|x_{k+1} - x_k\|_1 \leq \tau$$

- Bound becomes more stringent as n grows

New: $$\|x_{k+1} - x_k\|_1 \leq n\tau$$

- Reduces iteration count by 50%
- Disagreements in ranking $\leq 1.5\%$

Properties of Power Method

- Converges to unique vector
- Convergence rate α
- Vectorizes
- Storage for only a single vector
- Matrix vector multiplication with very sparse matrix
- Accurate (no subtractions)
- Simple (few decisions)

But: can be slow

PageRank Computation

- Power method
  Page, Brin, Motwani & Winograd 1999
- Acceleration of power method
  Brezinski & Redivo-Zaglia 2004
  Kamvar, Haveliwala, Manning & Golub 2003
  Haveliwala, Kamvar, Klein, Manning & Golub 2003
- 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

PageRank from Linear System

Eigenvector problem:

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

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]

Stationary Iterative Methods

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

- Can be faster than power method
- Can be faster than Krylov space methods
- Predictable, monotonic convergence
- Can converge even for $\alpha \approx 1$
- No failure due to "memory overload" (unlike Krylov space methods)
- Accurate (no subtractions)

[Gleich, Zhulov & Berkhin 2005]

Example [Gleich, Zhukov & Berkhin 2005]

Web graph: 1.4 billion nodes
6.6 billion edges
Beowulf cluster with 140 processors
Stopping criterion: residual norm $\leq 10^{-7}$

BiCGSTAB: 28.2 minutes (preconditioner?)

Power method: 35.5 minutes

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: $\quad D = I \quad O = \alpha S$
  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]
  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 $\quad \delta^T(I - \alpha H) =$ multiple of v^T

Then δ is multiple of π

[Arasu, Novak, Tomkins & Tomlin 2002]

Extension to Arbitrary w

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

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

1. Solve $\quad \delta^T = (1-\alpha)v^T(I - \alpha H)^{-1}$
2. Update $\quad \pi^T = \delta^T +$ stuff

This requires only two sparse solves

1. Solve: $\delta^T = (1-\alpha)v^T(I-\alpha H)^{-1}$

After similarity permutation:

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

$$\begin{pmatrix} \delta_1^T & \delta_2^T \end{pmatrix} \begin{pmatrix} I - \alpha H_1 & -\alpha H_2 \\ 0 & I \end{pmatrix} = (1-\alpha) \begin{pmatrix} v_1^T & v_2^T \end{pmatrix}$$

1. Sparse solve $\delta_1^T = (1-\alpha)v_1^T(I-\alpha H_1)^{-1}$
2. Set $\delta_2^T = \alpha\, \delta_1^T H_2 + (1-\alpha)v_2^T$

2. Update: $\pi^T = \delta^T +$ stuff

$$H = \begin{pmatrix} H_1 & H_2 \\ 0 & 0 \end{pmatrix} \qquad dw^T = \begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix} \begin{pmatrix} w_1^T & w_2^T \end{pmatrix}$$

1. Sparse solve $\quad z^T = \alpha w_1^T (I - \alpha H_1)^{-1}$
2. Set $\quad y^T = \alpha(w_2^T + z^T H_2)$
3. Update $\quad \pi^T = \delta^T + \underbrace{\frac{\|\delta_2\|_1}{1 - \|y\|_1} \begin{pmatrix} z^T & y^T \end{pmatrix}}_{\text{stuff}}$

PageRank via Linear System

$$\pi^T(I-\alpha S) = (1-\alpha)v^T \quad S = \begin{pmatrix} H_1 & H_2 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix} w^T$$

Arbitrary dangling node vector w, $w \geq 0$, $\|w\|_1 = 1$

Cost:

Two sparse solves with $I - \alpha H_1$ via Jacobi
Two matrix vector multiplications with H_2
Inner products and vector additions

Summary

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

- PageRank = left eigenvector of G
- PageRank insensitive to perturbations in G
- Adding in-links can only increase PageRank
- Adding out-links may decrease PageRank
- Improved stopping criterion
- Compute PageRank from linear system
- Jacobi can be competitive with Krylov methods
- Efficient treatment of dangling nodes