

Simplifying PageRank

Douglas V. de Jager

ICTIR 2009
Microsoft Research, Cambridge, UK

10 September, 2009

Accelerating Web growth

- July 2008
 - Google announced discovery of 10^{12} unique Web pages
 - Only a fraction of the discoverable Web
 - Google confessed to ranking less than 10^{11} discovered pages

Accelerating Web growth

- July 2008
 - Google announced discovery of 10^{12} unique Web pages
 - Only a fraction of the discoverable Web
 - Google confessed to ranking less than 10^{11} discovered pages
- What does tomorrow hold?
 - $> 10^9$ unique new web pages discovered / day in July 2008
 - 5 orders of magnitude increase in discovered Web in a decade
 - Rate of Web growth is ever increasing

Accelerating Web growth

- July 2008
 - Google announced discovery of 10^{12} unique Web pages
 - Only a fraction of the discoverable Web
 - Google confessed to ranking less than 10^{11} discovered pages
- What does tomorrow hold?
 - $> 10^9$ unique new web pages discovered / day in July 2008
 - 5 orders of magnitude increase in discovered Web in a decade
 - Rate of Web growth is ever increasing
- What about PageRank variants?
 - TrustRank

Large, dense matrix

Limiting probability of a Random Surfer being at a Web page

Large, dense matrix

Limiting probability of a Random Surfer being at a Web page

Random Surfer follows hyperlinks

$$\vec{p}^{(k)} \mathbf{H} = \vec{p}^{(k+1)},$$

$$H_{ij} = \begin{cases} \frac{1}{\deg(h_i)} & : \text{if } i \text{ links to } j \\ 0 & : \text{otherwise} \end{cases}$$

Large, dense matrix

Limiting probability of Random Surfer being at a Web page

When at dangling pages

$$\vec{p}^{(k)}(\mathbf{H} + \vec{d}\vec{v}^T) = \vec{p}^{(k+1)},$$

column vectors \vec{d}, \vec{v} are such that

$$d_i = \begin{cases} 1 & : \text{if } \deg(u_i) = 0 \\ 0 & : \text{otherwise} \end{cases}$$

and \vec{v} is a probability vector

Large, dense matrix

Limiting probability of Random Surfer being at a Web page

When bored

$$\vec{p}^{(k)}(c(\mathbf{H} + \vec{d}\vec{v}^T) + (1-c)\vec{1}\vec{v}^T) = \vec{p}^{(k+1)},$$

$$0 < c < 1$$

Large, dense matrix

Standard PageRank definition

$$\vec{\pi}(c(\mathbf{H} + \vec{d}\vec{v}^T) + (1 - c)\vec{1}\vec{v}^T) = \vec{\pi}$$

Large, dense matrix

Standard PageRank definition

$$\vec{\pi}(c(\mathbf{H} + \vec{d}\vec{v}^T) + (1 - c)\vec{1}\vec{v}^T) = \vec{\pi}$$

Sparse formulation of PageRank

$$c\vec{\pi}(\mathbf{H} + \vec{d}\vec{v}^T) + (1 - c)\vec{v} = \vec{\pi}$$

Cost of dangling pages

Optimised sparse algorithm

$$c\vec{\pi}^{(n)}\mathbf{H} + (1 - c + c \sum_{i \text{ is dangling}} \pi_i^{(n)})\vec{v} = \vec{\pi}^{(n+1)}$$

Cost of dangling pages

Optimised sparse algorithm

$$c\vec{\pi}^{(n)}\mathbf{H} + (1 - c + c \sum_{i \text{ is dangling}} \pi_i^{(n)})\vec{v} = \vec{\pi}^{(n+1)}$$

Percentage of dangling pages appears to be increasing

Cost of dangling pages

Optimised sparse algorithm

$$c\vec{\pi}^{(n)}\mathbf{H} + (1 - c + c \sum_{i \text{ is dangling}} \pi_i^{(n)})\vec{v} = \vec{\pi}^{(n+1)}$$

Time to sum dangling $\pi_i^{(n)}$ across tree network

$$\Theta\left(\frac{\text{dangling pages}}{\text{processors}} + \log(\text{processors})\right)$$

Cost of dangling pages

Optimised sparse algorithm

$$c\vec{\pi}^{(n)}\mathbf{H} + (1 - c + c \sum_{i \text{ is dangling}} \pi_i^{(n)})\vec{V} = \vec{\pi}^{(n+1)}$$

Cost of summing dangling $\pi_i^{(n)}$

$$\Theta(\text{dangling pages} + \text{processors} \times \log(\text{processors}))$$

Cost of dangling pages

Optimised sparse algorithm

$$c\vec{\pi}^{(n)}\mathbf{H} + (1 - c + c \sum_{i \text{ is dangling}} \pi_i^{(n)})\vec{V} = \vec{\pi}^{(n+1)}$$

Messaging cost of summing dangling $\pi_i^{(n)}$:

$$\Theta(\text{processors} \times \log(\text{processors}))$$

Other tractability research

- Hypergraph partitioning ×
- Asynchronous solution algorithms ×
- No reduction in dangling-page messaging

No dangling-page matrix

Scalar multiple of PageRank

$$c\vec{\phi}\mathbf{H} + \vec{v} = \vec{\phi}$$

No dangling-page matrix

Scalar multiple of PageRank

$$c\vec{\phi}\mathbf{H} + \vec{v} = \vec{\phi}$$

- Collection of proofs
 - New proof for each PageRank variant
 - Lumpability theory; Sherman–Morrison formula; *ad hoc* reformulation

No dangling-page matrix

Scalar multiple of PageRank

$$c\vec{\phi}\mathbf{H} + \vec{v} = \vec{\phi}$$

- Collection of proofs
 - New proof for each PageRank variant
 - Lumpability theory; Sherman–Morrison formula; *ad hoc* reformulation
 - Our aim is to provide a single, accessible means of proof

Recursively removing dangling pages

Non-dangling page, n

$$\phi_n = c \sum_{(j \text{ is non-dangling})} \phi_j H_{jn} + v_n$$

Dangling page, d

$$\phi_d = c \sum_{(j \text{ is non-dangling})} \phi_j H_{jd} + v_d$$

Recursively removing dangling pages

Non-dangling page, n

$$\phi_n = c \sum_{(j \text{ is non-dangling})} \phi_j H_{jn} + v_n$$

Dangling page, d

$$\phi_d = c \sum_{(j \text{ is non-dangling})} \phi_j H_{jd} + v_d$$

- Recursively remove dangling pages
 - Dangling pages have no update role
 - Removing dangling pages yields more dangling pages

Examples to motivate main theorem beyond Page

- Two problems which share a solution vector:

$$\bullet \vec{x} = \begin{pmatrix} 0 & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x}$$

$$\bullet \vec{x} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x} + \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ 0 \end{pmatrix}$$

Examples to motivate main theorem beyond Page

- Two problems which share a solution vector:

- $\vec{x} = \begin{pmatrix} 0 & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x}$

- $\vec{x} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x} + \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ 0 \end{pmatrix}$

- Pleasant properties of second problem:

- Finite termination: $x_1 = \frac{1}{3}$; $x_3 = x_1 = \frac{1}{3}$; $x_2 = x_3 + \frac{2}{3} = 1$
- Sparsity
- Spectral radius
- Asynchronous algorithms applicable

Examples to motivate main theorem beyond PageRank

- How many solutions?

- $\vec{x} = \begin{pmatrix} 0 & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x}$

Examples to motivate main theorem beyond PageRank

- How many solutions?

- $\vec{x} = \begin{pmatrix} 0 & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x}$

✓ one solution

Examples to motivate main theorem beyond PageRank

- How many solutions?

- $\vec{x} = \begin{pmatrix} 0 & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x}$

✓ one solution

- $\vec{x} = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \vec{x}$

Examples to motivate main theorem beyond PageRank

- How many solutions?

- $\vec{x} = \begin{pmatrix} 0 & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x}$

✓ one solution

- $\vec{x} = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \vec{x}$

✗ no solution

Examples to motivate main theorem beyond PageRank

- How many solutions?

- $\vec{x} = \begin{pmatrix} 0 & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x}$

✓ one solution

- $\vec{x} = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \vec{x}$

✗ no solution

- $\vec{x} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \vec{x}$

Examples to motivate main theorem beyond PageRank

- How many solutions?

- $\vec{x} = \begin{pmatrix} 0 & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & 1 \\ 1 & 0 & 0 \end{pmatrix} \vec{x}$

✓ one solution

- $\vec{x} = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \vec{x}$

× no solution

- $\vec{x} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \vec{x}$

✓ one solution

Main Theorem

Suppose:

- $\mathbf{M} \in \mathbb{C}^{n \times n}$; column $\vec{a}, \vec{b} \in \mathbb{C}^n$, where $\forall i, b_i = 0$ or 1 .

Main Theorem

Suppose:

- $\mathbf{M} \in \mathbb{C}^{n \times n}$; column $\vec{a}, \vec{b} \in \mathbb{C}^n$, where $\forall i, b_i = 0$ or 1 .
- 1 is **not** an eigenvalue of $(\mathbf{M} - \vec{b}\vec{a}^T)$.

Main Theorem

Suppose:

- $\mathbf{M} \in \mathbb{C}^{n \times n}$; column $\vec{a}, \vec{b} \in \mathbb{C}^n$, where $\forall i, b_i = 0$ or 1 .
- 1 is **not** an eigenvalue of $(\mathbf{M} - \vec{b}\vec{a}^T)$.

Then:

If 1 is an eigenvalue of \mathbf{M} , then

- It is a simple eigenvalue
- A corresponding right eigenvector, \vec{x} , of \mathbf{M} uniquely satisfies:

$$\vec{x} = \vec{x}(\mathbf{M} - \vec{b}\vec{a}^T) + \vec{a}$$

Using the main theorem

Yields scalar multiple of PageRank

$$\vec{x} = c\vec{x}\mathbf{H} + (1 - c)\vec{x}\vec{1}\vec{v}^T + c\vec{v} = c\vec{x}\mathbf{H} + e\vec{v},$$

where e is a non-zero scalar,
as $\sum_i x_i$ is neither 0 nor $-c$.

NB e serves only to scale \vec{x} .

Using the main theorem

Yields scalar multiple of PageRank

$$\vec{x} = c\vec{x}\mathbf{H} + (1 - c)\vec{x}\vec{1}\vec{v}^T + c\vec{v} = c\vec{x}\mathbf{H} + e\vec{v},$$

where e is a non-zero scalar,
as $\sum_i x_i$ is neither 0 nor $-c$.

NB e serves only to scale \vec{x} .

- TrustRank is similar, though two equations are produced.

Is PageRank still the bottleneck?

- **Suppose:**
 - Only new or changed pages need indexing
 - New or changed pages per day \rightarrow pages as pages $\rightarrow \infty$

Is PageRank still the bottleneck?

- **Suppose:**
 - Only new or changed pages need indexing
 - New or changed pages per day \rightarrow pages as pages $\rightarrow \infty$
- **Then**
 - Indexing $< \Theta$ (pages)

Is PageRank still the bottleneck?

- Suppose:
 - Only new or changed pages need indexing
 - New or changed pages per day \rightarrow pages as pages $\rightarrow \infty$
- Then
 - Indexing $< \Theta$ (pages)
- To be determined
 - Recursively reduced PageRank $< \Theta$ (pages)?

Thank you!