Outline

1. Elements of Web Search [Bryan and Leise, 2006, Gleich, 2015]
2. PageRank [Bryan and Leise, 2006]
3. Google PageRank and Beyond [Langville and Meyer, 2006]
4. Readings

Elements of Web Search [Bryan and Leise, 2006, Gleich, 2015]

Elements of Web Search

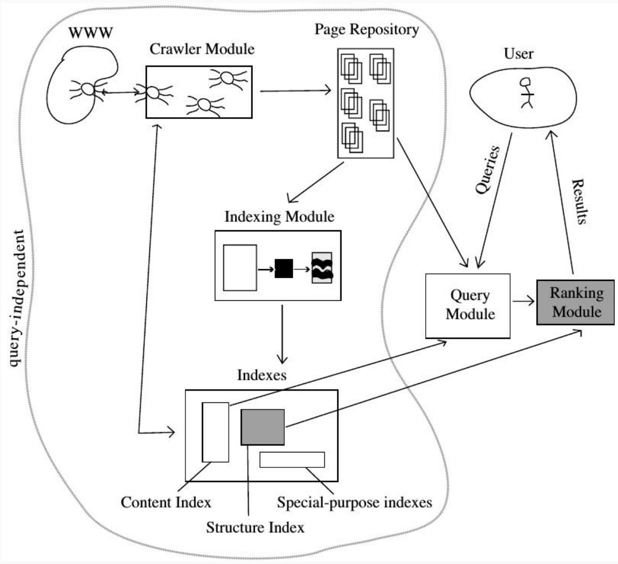


Figure 1: *Google's PageRank and Beyond*, Langville and Meyer

Term Document Matrices

- Start with dictionary of terms
- Index each document - Count f_{ij}, # times term i appears in document j
- Term Document Matrix

Vector Space Model

- Document vector and Query vector
- Similarity Scores
- Dumais's Improvement - Latent Semantic Indexing[1]

[1] http://www2.denizyuret.com/ref/berry/berry95using.pdf

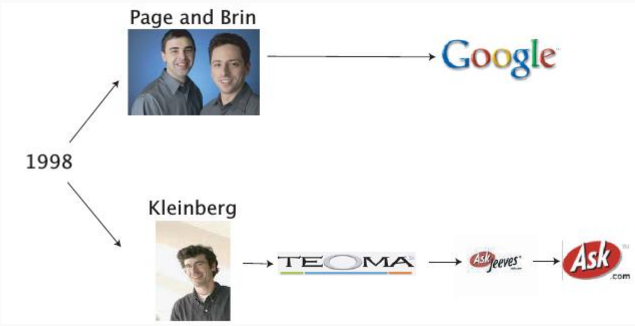
Web IR??

- It is HUGE
  - Over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web
- It is DYNAMIC
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year
- It is SELF-ORGANIZED
  - no standards, review process, formats
  - errors, falsehoods, link rot, and spammers

"It is HYPERLINKED!"

PageRank [Bryan and Leise, 2006]

Link Analysis[2]



[2] The book by Barabasi, *Linked: The New Science of Networks* : learning valuable information about networks ranging from the AIDS transmission and power grid networks to terrorists and email networks.

Eigen Vectors?

- HITS
- Google Pagerank
- Eigenvector computation: 2×2 matrix example
- A village full of ethical thieves
- Power method (Lancsoz)

The $25,000,000,000 GOOGLE[4]

- Approximate market value of GOOGLE when it went public in 2004

[3] http://toolbar.google.com

[4] http://www.google.com/technology/index.html - The heart of Google's software is Page rank

The $25,000,000,000 GOOGLE[4]

- Approximate market value of GOOGLE when it went public in 2004
- PageRank Score[3]

[3] http://toolbar.google.com
[4] http://www.google.com/technology/index.html - The heart of Google's software is Page rank

PageRank

- Assign some measure of importance to every web page based on endorsements

PageRank

- Assign some measure of importance to every web page based on endorsements
- Web as a directed graph $G = (V, E)$, (v_j, v_i) is an edge of E if page v_j has a link to page v_i

PageRank

- Assign some measure of importance to every web page based on endorsements
- Web as a directed graph $G = (V, E)$, (v_j, v_i) is an edge of E if page v_j has a link to page v_i
- Rank of a page is the sum of the ranks of the pages that point to it, divided by their degrees

$$r_i = \sum_{j:(v_j,v_i)\in E} \frac{r_j}{d_{out}(v_j)} \tag{1}$$

Example

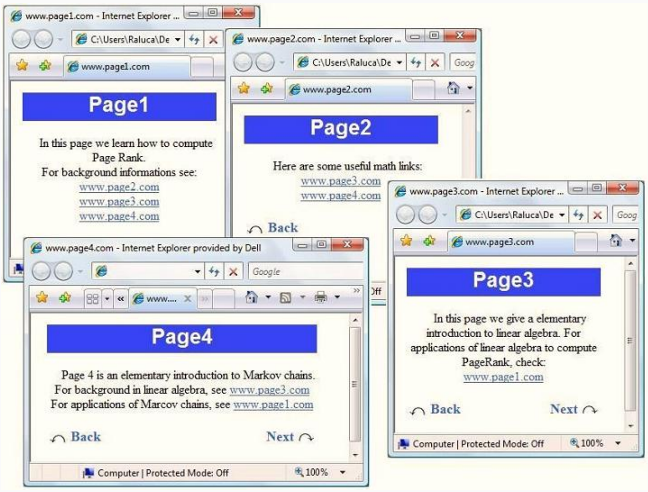


Figure 2: `http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html`

Example : Continued[5]

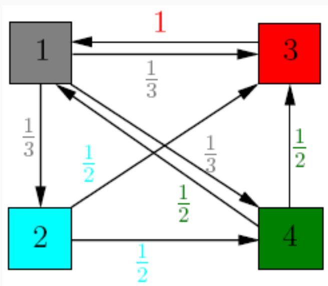


Figure 3: Graph Model

$$
\begin{aligned}
r_1 &= r_3 + \frac{1}{2} r_4 \\
r_2 &= \frac{1}{3} r_1 \\
r_3 &= \frac{1}{3} r_1 + \frac{1}{2} r_2 + \frac{1}{2} r_4 \\
r_4 &= \frac{1}{3} r_1 + \frac{1}{2} r_2
\end{aligned} \tag{2}
$$

[5]Eigen vector problem $Hr = 1.r$, i.e. find an eigenvector of H corresponding to eigenvalue 1 where $H = AD^{-1}$, A is the adjacency matrix and D is the diagonal matrix with the out-degrees of the nodes on the diagonal.

Random Surfer

- Think of a random surfer on the web browsing/travelling pages/states

[6] $M := [m_{ij}]$ is a column stochastic Markov matrix.

Random Surfer

- Think of a random surfer on the web browsing/travelling pages/states
- Let the transition of a surfer from state j to state i be guided by transition probability m_{ij}[6]

$$H = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \quad (3)$$

[6] $M := [m_{ij}]$ is a column stochastic Markov matrix.

Random Surfer

- Think of a random surfer on the web browsing/travelling pages/states
- Let the transition of a surfer from state j to state i be guided by transition probability m_{ij}[6]

$$H = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \tag{3}$$

- What will be the stationary probability distribution on four states of this Markov chain?

$$\lim_{k \to \infty} H^k r \tag{4}$$

where r is any arbitrary probability distribution on states.

[6] $M := [m_{ij}]$ is a column stochastic Markov matrix.

Google PageRank and Beyond [Langville and Meyer, 2006]

Difficulties and Teleportation!

- Non-unique rankings

Difficulties and Teleportation!

- Non-unique rankings
- Dangling nodes

Difficulties and Teleportation!

- Non-unique rankings
- Dangling nodes
- Solution:

$$S = H + \frac{1}{n} e a^T, \;\; a_i = 1 \; \textit{if} \;\; i \;\textit{is a dangling node} \tag{5}$$

$$G = \alpha S + \frac{1}{n}(1 - \alpha) e e^T \tag{6}$$

Difficulties and Teleportation!

- Non-unique rankings
- Dangling nodes
- Solution:

$$S = H + \frac{1}{n}ea^T, \; a_i = 1 \; if \; i \; is \; a \; dangling \; node \tag{5}$$

$$G = \alpha S + \frac{1}{n}(1-\alpha)ee^T \tag{6}$$

- G is called the **Google Matrix**

Difficulties and Teleportation!

- Non-unique rankings
- Dangling nodes
- Solution:

$$S = H + \frac{1}{n} e a^T, \; a_i = 1 \; if \; i \; is \; a \; dangling \; node \tag{5}$$

$$G = \alpha S + \frac{1}{n}(1 - \alpha) e e^T \tag{6}$$

- G is called the **Google Matrix**
- G is column stochastic

Analysis

1. Every column stochastic matrix has 1 as an eigenvalue.
2. If a matrix is positive and column stochastic, then any eigenvector in V_1 has all positive or all negative components.
3. Let v and w be linearly independent vectors in $\mathbb{R}^m$, $m \geq 2$. Then, for some values of s and t that are not both zero, the vector $x = sv + tw$ has both positive and negative components.
4. If a matrix is positive and column stochastic, then V_1 has dimension 1.

Readings

References

Bryan, K. and Leise, T. (2006).
The $25,000,000,000 eigenvector: The linear algebra behind google.
SIAM Review, 48(3):569–581.

Gleich, D. F. (2015).
Pagerank beyond the web.
SIAM Review, 57(3):321–363.

Langville, A. and Meyer, C. (2006).
Google's PageRank and Beyond: The Science of Search Engine Rankings.
Princeton University Press.

Questions?