Introduction to the use of Link Analysis by Web Search Engines

Amy Langville
Carl Meyer

Department of Mathematics
North Carolina State University
Raleigh, NC
Outline

• Introduction to Information Retrieval (IR)

• Link Analysis

• HITS Algorithm

• PageRank Algorithm
Short History of IR

IR = search within doc. coll. for particular info. need (query)

B. C. cave paintings
7-8th cent. A.D. Beowulf
12th cent. A.D. invention of paper, monks in scriptoriums
1450 Gutenberg’s printing press
1700s Franklin’s public libraries
1872 Dewey’s decimal system
Card catalog
1940s-1950s Computer
1960s Salton’s SMART system
1989 Berner-Lee’s WWW
the pre-1998 Web

Yahoo
- hierarchies of sites
- organized by humans

Best Search Techniques
- word of mouth
- expert advice

Overall Feeling of Users
- Jorge Luis Borges’ 1941 short story, *The Library of Babel*

When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness. All men felt themselves to be the masters of an intact and secret treasure. There was no personal or world problem whose eloquent solution did not exist in some hexagon.

... As was natural, this inordinate hope was followed by an excessive depression. The certitude that some shelf in some hexagon held precious books and that these precious books were inaccessible, seemed almost intolerable.
1998 ... enter Link Analysis

Change in User Attitudes about Web Search

Today

- “It’s not my homepage, but it might as well be. I use it to ego-surf. I use it to read the news. Anytime I want to find out anything, I use it.” - Matt Groening, creator and executive producer, The Simpsons

- “I can’t imagine life without Google News. Thousands of sources from around the world ensure anyone with an Internet connection can stay informed. The diversity of viewpoints available is staggering.” - Michael Powell, chair, Federal Communications Commission

- “Google is my rapid-response research assistant. On the run-up to a deadline, I may use it to check the spelling of a foreign name, to acquire an image of a particular piece of military hardware, to find the exact quote of a public figure, check a stat, translate a phrase, or research the background of a particular corporation. It’s the Swiss Army knife of information retrieval.” - Garry Trudeau, cartoonist and creator, Doonesbury
Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR
Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?
Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It’s huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 550 billion pages
Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It’s huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 550 billion pages

- It’s dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year
Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It’s huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 550 billion pages

- It’s dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year

- It’s self-organized.
  - no standards, review process, formats
  - errors, falsehoods, link rot, and spammers!
Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

• It’s huge.
  – over 10 billion pages, average page size of 500KB
  – 20 times size of Library of Congress print collection
  – Deep Web - 550 billion pages

• It’s dynamic.
  – content changes: 40% of pages change in a week, 23% of .com change daily
  – size changes: billions of pages added each year

• It’s self-organized.
  – no standards, review process, formats
  – errors, falsehoods, link rot, and spammers!

A Herculean Task!
Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It’s huge.
  - over 10 billion pages, each about 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 550 billion pages

- It’s dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year

- It’s self-organized.
  - no standards, review process, formats
  - errors, falsehoods, link rot, and spammers!

- Ah, but it’s **hyperlinked**!
  - Vannevar Bush’s 1945 memex
Elements of a Web Search Engine

- WWW
- Crawler Module
- Indexing Module
- Page Repository
- User
- Indexes
  - Content Index
  - Structure Index
  - Special-purpose indexes
- Query Module
- Ranking Module
Indexing Wars

![Graph showing the growth of different search engines from 12/95 to 12/02](image)

**Actual Index King =**

Internet Archive - [http://web.archive.org](http://web.archive.org)
Query Processing

Step 1: User enters query, i.e., aztec baby

Step 2: Inverted file consulted

- term 1 (aardvark) - 3, 117, 3961
- term 10 (aztec) - 3, 15, 19, 101, 673, 1199
- term 11 (baby) - 3, 31, 56, 94, 673, 909, 11114, 253791
- term m (zymurgy) - 1159223

Step 3: Relevant set identified, i.e. (3, 673)

Simple traditional engines stop here.
Modification to Inverted File

- add more features to inverted file by appending vector to each page identifier, i.e., [in title?, in descrip.?, # of occurrences]

- Modified inverted file

  - term 1 (aardvark) - 3 [0,0,3], 117 [1,1,10], 3961 [0,1,4]
  - term 10 (aztec) - 3 [1, 1, 27], 15 [0,0,1], 19 [1,1,21], 101 [0,1,7], 673 [0, 0, 3], 1199 [0,0,3]
  - term 11 (baby) - 3 [1, 1, 10], 31 [0,0,2], 56 [0,1,3], 94 [1,1,11], 673 [1, 1, 14], 909 [0,0,2], 11114 [1,1,22], 253791 [0,1,6]
  - term m (zymurgy) - 1159223 [1,1,9]

- IR score computed for each page in relevant set.

  EX: IR score (page 3) = \((1 + 1 + 27) \times (1 + 1 + 10) = 348\)
  IR score (page 673) = \((0 + 0 + 3) \times (1 + 1 + 14) = 48\)

  Early web engines stop here.

  Problem = Ranking by IR score is not good enough.
CSC issues in Crawling and Indexing

- create parallel crawlers but avoid overlap
- ethical spidering
- how often to crawl pages, which pages to update
- best way to store huge inverted file
- how to efficiently update inverted file
- store the files across processors
- provide for parallel access
- create robust, failure-resistant system
Link Analysis

- uses hyperlink structure to focus the relevant set
- combine IR score with popularity or importance score

PageRank - Brin and Page

HITS - Kleinberg
The Web as a Graph

Nodes = webpages
Arcs = hyperlinks
Web Graphs

CSC and MATH problems here:

- store adjacency matrix
- update adjacency matrix
- visualize web graph
- locate clusters in graph
How to Use Web Graph for Search

Hyperlink = Recommendation

- page with 20 recommendations (inlinks) must be more important than page with 2 inlinks.
- but status of recommender matters.
  EX: letters of recommendation: 1 letter from Trump vs. 20 from unknown people
- but what if recommender is generous with recommendations?
  EX: suppose Trump has written over 40,000 letters.
- each inlink should be weighted to account for status of recommender and # of outlinks from that recommender
How to Use Web Graph for Search

Hyperlink = Recommendation

- page with 20 recommendations (inlinks) must be more important than page with 2 inlinks.
- but status of recommender matters.
  EX: letters of recommendation: 1 letter from Trump vs. 20 from unknown people
- but what if recommender is generous with recommendations?
  EX: suppose Trump has written over 40,000 letters.
- each inlink should be weighted to account for status of recommender and # of outlinks from that recommender

PAGERANK - importance/popularity score given to each page
Our Search: Google Technology

Google searches more sites more quickly, delivering the most relevant results.

Introduction

Google runs on a unique combination of advanced hardware and software. The speed you experience can be attributed in part to the efficiency of our search algorithm and partly to the thousands of low cost PC's we've networked together to create a superfast search engine.

The heart of our software is PageRank™, a system for ranking web pages developed by our founders Larry Page and Sergey Brin at Stanford University. And while we have dozens of engineers working to improve every aspect of Google on a daily basis, PageRank continues to provide the basis for all of our web search tools.

PageRank Explained

PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."

Important, high-quality sites receive a higher PageRank, which Google remembers each time it conducts a search. Of course, important pages mean nothing to you if they don't match your query. So, Google combines PageRank with sophisticated text-matching techniques to find pages that are both important and relevant to your search. Google goes far beyond the number of times a term appears on a page and examines all aspects of the page's content (and the content of the pages linking to it) to determine if it's a good match for your query.

Integrity

Google's complex, automated methods make human tampering with our results extremely difficult. And though we do run relevant ads above and next to our results, Google does not sell placement within the results themselves (i.e., no one can buy a higher PageRank). A Google search is an easy, honest and objective way to find high-quality websites with information relevant to your search.
Another Way to Use Web Graph for Search

- give each page 2 scores (hub and authority scores) instead of just 1.

- **DEFN:**
  - **Authorities**
  - **Hubs**

- pages can be both hubs and authorities (EX: ATL airport)

- Good hub pages point to good authority pages, and good authorities are pointed to by good hubs.

**HITS** - hub and authority score given to each page

**HITS** - (Hypertext Induced Topic Search)
Pop Quiz

$$\mathbf{Ax}_i = \lambda_i \mathbf{x}_i$$

$$\sigma = [\lambda_1, \lambda_2, \ldots, \lambda_n]$$

Classic Method to Compute Dominant Eigenpair?
Pop Quiz

\[ \mathbf{A} \mathbf{x}_i = \lambda_i \mathbf{x}_i \]

\[ \sigma = [\lambda_1, \lambda_2, \ldots, \lambda_n] \]

Classic Method to Compute Dominant Eigenpair?

Power Method
Pop Quiz

\[ \mathbf{Ax}_i = \lambda_i \mathbf{x}_i \]

\[ \sigma = [\lambda_1, \lambda_2, \ldots, \lambda_n] \]

**Classic Method to Compute Dominant Eigenpair?**

**Power Method**

- The Iterative Procedure
  \[ \mathbf{x}^{(k)} = \mathbf{A}\mathbf{x}^{(k-1)} \]
  converges to \( \mathbf{x}_1 \) for \( |\lambda_1| > |\lambda_2| \) regardless of the starting vector \( \mathbf{x}^{(0)} \).

- Rate of Convergence: rate at which \( (|\lambda_2|/|\lambda_1|)^k \rightarrow 0 \).

- Notice \( \mathbf{x}^{(k)} = \mathbf{A}^k \mathbf{x}^{(0)} \).
HITS Algorithm
Hypertext Induced Topic Search (J. Kleinberg 1998)

Determine Authority & Hub Scores

- $a_i$ = authority score for $P_i$
- $h_i$ = hub score for $P_i$

Successive Refinement

- Start with $h_i(0) = 1$ for all pages $P_i$
- Successively refine rankings
  - For $k = 1, 2, \ldots$
    
    \[
    a_i(k) = \sum_{j: P_j \rightarrow P_i} h_j(k - 1) \quad \Rightarrow \quad a_k = L^T h_{k-1}
    \]
    
    \[
    h_i(k) = \sum_{j: P_i \rightarrow P_j} a_j(k) \quad \Rightarrow \quad h_k = L a_k
    \]

- $A = L^T L \quad a_k = A a_{k-1} \rightarrow \text{e-vector}$
- $H = LL^T \quad h_k = H h_{k-1} \rightarrow \text{e-vector}$
HITS Neighborhood Graph

1. Find relevant set by consulting inverted file
2. Build neighborhood graph
3. Compute authority & hub scores for just the neighborhood
HITS Example

1. Relevant set = [1, 6]

2. Neighborhood graph \( \Lambda \)

3. Compute authority & hub scores.

Adjacency matrix for \( N = L = \)

\[
\begin{pmatrix}
1 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{pmatrix}
\]
HITS Example (cont.)

Authority matrix \( A = L^T L \)

\[
L^T L = \begin{pmatrix}
1 & 2 & 3 & 5 & 6 & 10 \\
1 & 1 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 0 & 2 & 1 & 1 \\
5 & 0 & 0 & 1 & 1 & 0 \\
6 & 0 & 0 & 1 & 0 & 3 \\
10 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

Hub matrix \( H = LL^T \)

\[
LL^T = \begin{pmatrix}
1 & 2 & 3 & 5 & 6 & 10 \\
1 & 2 & 0 & 1 & 0 & 1 & 1 \\
2 & 0 & 1 & 0 & 0 & 0 & 0 \\
3 & 1 & 0 & 1 & 0 & 0 & 1 \\
5 & 0 & 0 & 0 & 0 & 0 & 0 \\
6 & 1 & 0 & 0 & 0 & 2 & 0 \\
10 & 1 & 0 & 1 & 0 & 0 & 1
\end{pmatrix}
\]

Authority score vector \( \mathbf{a} \)

\[
\mathbf{a}^T = \begin{pmatrix}
1 & 2 & 3 & 5 & 6 & 10 \\
0 & 0 & .3660 & .1340 & .5 & 0
\end{pmatrix}
\]

Hub score vector \( \mathbf{h} \)

\[
\mathbf{h}^T = \begin{pmatrix}
1 & 2 & 3 & 5 & 6 & 10 \\
.3660 & 0 & .2113 & 0 & .2113 & .2113
\end{pmatrix}
\]
HITS Convergence

- HITS with normalization step always converges.
- Rate of convergence depends on eigengap $\lambda_1 - \lambda_2$.
- BUT $\lambda_1$ may be a repeated root $\Rightarrow$ nonunique solutions. Different $h_0$ and $a_0$ can lead to different $h_\infty$ and $a_\infty$.
- $h_\infty$ and $a_\infty$ can contain 0 values for some pages, which is undesirable in ranking context.
Pros & Cons

Advantages

- Returns satisfactory results
  - Client gets both authority & hub scores
- Some flexibility

Disadvantages

- Too much must happen while client is waiting; query-dependent
  - Custom built neighborhood graph needed for each query
  - Two eigenvector computations needed for each query
- Scores can be manipulated by creating artificial hubs

Modified HITS in Teoma
CSC and MATH Issues with HITS

- how to form $N$ and fix topic drift problem
- incorporating weights into $L$ matrix
- fast eigenvector computation, beating the power method
- updating $L$, $h$, and $a$ for query-independent HITS
Markov Chain

- Transition Matrix $P$ is ?
Pop Quiz

Markov Chain

- Transition Matrix $P$ is square and stochastic

EX: $P = \begin{pmatrix} .7 & .3 \\ .45 & .55 \end{pmatrix}$
Pop Quiz

Markov Chain

- Transition Matrix $P$ is?
  - square and stochastic
  
  EX: $P = \begin{pmatrix} S & R \\ S & 0.7 & 0.3 \\ R & 0.45 & 0.55 \end{pmatrix}$

- Element $p_{ij}$ represents?
Pop Quiz

Markov Chain

- Transition Matrix $P$ is?
  
  - square and stochastic

- Element $p_{ij}$ represents?
  
  - probability of transitioning from state $i$ to state $j$

EX: $P = \begin{pmatrix} S & R \\ S & .7 & .3 \\ R & .45 & .55 \end{pmatrix}$
## Pop Quiz

### Markov Chain

- **Transition Matrix** \( P \) is ?
  
  square and stochastic

  \[
  \begin{pmatrix}
  S & R \\
  .7 & .3 \\
  .45 & .55 
  \end{pmatrix}
  \]

- **Element** \( p_{ij} \) represents ?
  
  probability of transitioning from state \( i \) to state \( j \)

- **Stationary distribution** \( \pi^T \) satisfies ?
Pop Quiz

Markov Chain

• Transition Matrix $P$ is?
  
  square and stochastic

  EX: $P = \begin{pmatrix} S & R \\ S & .7 & .3 \\ R & .45 & .55 \end{pmatrix}$

• Element $p_{ij}$ represents?
  
  probability of transitioning from state $i$ to state $j$

• Stationary distribution $\pi^T$ satisfies?
  
  $\pi^T = \pi^T P$
Markov Chain

- Transition Matrix $P$ is?
  - square and stochastic
  
  EX: $P = \begin{pmatrix} S & R \\ S & .7 \\ R & .45 \end{pmatrix}$

- Element $p_{ij}$ represents?
  - probability of transitioning from state $i$ to state $j$

- Stationary distribution $\pi^T$ satisfies?
  - $\pi^T = \pi^T P$

- $\pi_i$ represents?
Pop Quiz

Markov Chain

• Transition Matrix \( P \) is?

  \[ P = \begin{pmatrix} S & R \\ S & 0.7 \\ R & 0.3 \end{pmatrix} \]

  square and stochastic

• Element \( p_{ij} \) represents?

  probability of transitioning from state \( i \) to state \( j \)

• Stationary distribution \( \pi^T \) satisfies?

  \[ \pi^T = \pi^T P \]

• \( \pi_i \) represents?

  long-run proportion of time spent in state \( i \)
Pop Quiz

Markov Chain

- Transition Matrix $P$ is?
  - square and stochastic
  - $P = \begin{pmatrix} .7 & .3 \\ .45 & .55 \end{pmatrix}$

- Element $p_{ij}$ represents?
  - probability of transitioning from state $i$ to state $j$

- Stationary distribution $\pi^T$ satisfies?
  - $\pi^T = \pi^T P$

- $\pi_i$ represents?
  - long-run proportion of time spent in state $i$

- Classic Method for finding $\pi^T$
Pop Quiz

Markov Chain

- Transition Matrix $P$ is?
  
  square and stochastic

  EX: $P = \begin{pmatrix} S & R \\ R & S \end{pmatrix} = \begin{pmatrix} .7 & .3 \\ .45 & .55 \end{pmatrix}$

- Element $p_{ij}$ represents?
  
  probability of transitioning from state $i$ to state $j$

- Stationary distribution $\pi^T$ satisfies?

  $\pi^T = \pi^T P$

- $\pi_i$ represents?

  long-run proportion of time spent in state $i$

- Classic Method for finding $\pi^T$?

  Power Method, rate of convergence = rate at which $|\lambda_2|^k \rightarrow 0$
The PageRank Idea

• Ranking is preassigned

• Your page \( P \) has some rank \( r(P) \)

• Adjust \( r(P) \) higher or lower depending on ranks of pages that point to \( P \)

• Importance is not just number, but *quality* of in-links
  — role of outlinks relegated
  — much less sensitive to spamming
PageRank

The Definition

- \( r(P) = \sum_{P \in B_P} \frac{r(P)}{|P|} \)
  - \( B_P = \{\text{all pages pointing to } P\} \)
  - \(|P| = \text{number of out links from } P\)

Successive Refinement

- Start with \( r_0(P_i) = 1/n \) for all pages \( P_1, P_2, \ldots, P_n \)
- Iteratively refine rankings for each page

\[ r_1(P_i) = \sum_{P \in B_{P_i}} \frac{r_0(P)}{|P|} \]
\[ r_2(P_i) = \sum_{P \in B_{P_i}} \frac{r_1(P)}{|P|} \]
\[ \vdots \]
\[ r_{j+1}(P_i) = \sum_{P \in B_{P_i}} \frac{r_j(P)}{|P|} \]
In Matrix Notation

After Step $j$

$$\pi_j^T = [r_j(P_1), r_j(P_2), \cdots, r_j(P_n)]$$

$$\pi_{j+1}^T = \pi_j^T \mathbf{P} \quad \text{where} \quad p_{ij} = \begin{cases} \frac{1}{|P_i|} & \text{if } i \rightarrow j \\ 0 & \text{o.w.} \end{cases}$$
In Matrix Notation

After Step $j$

$$\pi_j^T = [r_j(P_1), r_j(P_2), \cdots, r_j(P_n)]$$

$$\pi_{j+1}^T = \pi_j^T \mathbf{P} \quad \text{where} \quad p_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{o.w.} \end{cases}$$

$$\mathbf{P} = \begin{pmatrix}
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}$$
In Matrix Notation

After Step $j$

$$\pi_j^T = [r_j(P_1), r_j(P_2), \cdots, r_j(P_n)]$$

$$\pi_{j+1}^T = \pi_j^T P$$ where $p_{ij} = \begin{cases} 1/|P_i| & \text{if } i \to j \\ 0 & \text{o.w.} \end{cases}$

\[
P = \begin{pmatrix}
p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\
p_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
p_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
p_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
p_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
p_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
p_6 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}
\]

PageRank = $\lim_{j \to \infty} \pi_j^T = \pi^T$ (provided limit exists)

It’s Almost a Markov Chain

$P$ has row sums = 1 for ND nodes, row sums = 0 for D nodes
In Matrix Notation

It’s Almost a Markov Chain

- $P$ has row sums = 1 for ND nodes, row sums = 0 for D nodes
In Matrix Notation

It’s Almost a Markov Chain

- $P$ has row sums = 1 for ND nodes, row sums = 0 for D nodes

Stochasticity Fix: $\tilde{P} = P + av^T$. $(a_i = 1$ for $i \in D$, 0, o.w.)
In Matrix Notation

It’s Almost a Markov Chain

- \( P \) has row sums = 1 for ND nodes, row sums = 0 for D nodes

**Stochasticity Fix:**  \( \tilde{P} = P + av^T \).

\[
\tilde{P} = \begin{bmatrix}
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}, \text{where } a = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, v^T = 1/6 \ e^T
\]
In Matrix Notation

It’s Almost a Markov Chain

- $P$ has row sums $= 1$ for ND nodes, row sums $= 0$ for D nodes

**Stochasticity Fix:** $\tilde{P} = P + av^T$.  \hspace{1cm} (a$_i$=1 for $i \in D$, 0, o.w.)

$$
\tilde{P} = \begin{bmatrix}
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}, \text{where } a= \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, v^T=1/6 e^T
$$

- Each $\pi_j^T$ is a probability distribution vector \hspace{1cm} ($\sum r_j(P_i)=1$)
- $\pi_{j+1}^T = \pi_j^T \tilde{P}$ is random walk on the graph defined by links
- $\pi^T = \lim_{j \to \infty} \pi_j^T$ = stationary probability distribution
Random Surfer

Web Surfer Randomly Clicks On Links (Back button not a link)

Long-run proportion of time on page $P_i$ is $\pi_i$

Problems
Random Surfer

Web Surfer Randomly Clicks On Links
    Long-run proportion of time on page $P_i$ is $\pi_i$

Problems
    Dead end page (nothing to click on)  ($\pi^T$ not well defined)
    Could get trapped into a cycle ($P_i \rightarrow P_j \rightarrow P_i$) (No convergence)
Random Surfer

Web Surfer Randomly Clicks On Links (Back button not a link)

Long-run proportion of time on page $P_i$ is $\pi_i$

Problems

Dead end page (nothing to click on) $(\pi^T$ not well defined)

Could get trapped into a cycle $(P_i \rightarrow P_j \rightarrow P_i)$ (No convergence)

Convergence

Markov chain must be irreducible and aperiodic
Random Surfer

Web Surfer Randomly Clicks On Links

Long-run proportion of time on page \( P_i \) is \( \pi_i \)

Problems

Dead end page (nothing to click on) \( (\pi^T \) not well defined) \( \)

Could get trapped into a cycle \( (P_i \rightarrow P_j \rightarrow P_i) \) (No convergence)

Convergence

Markov chain must be irreducible and aperiodic

DEFN: a chain is **irreducible** if every page is reachable from every other page.

DEFN: every **reducible** chain can be permuted to the form \[
\begin{bmatrix}
X & Y \\
0 & Z
\end{bmatrix}
\]
Bored Surfer Enters Random URL

Irreducibility Fix:  \[
\tilde{P} = \alpha \tilde{P} + (1 - \alpha)E \\
\tilde{P} = \alpha P + \alpha a v^T + (1 - \alpha)E
\]

- \( \pi^T \) is now guaranteed to exist and be unique and power method is guaranteed to converge to \( \pi^T \).
Random Surfer

Bored Surfer Enters Random URL

Irreducibility Fix: $\tilde{P} = \alpha P + (1 - \alpha)E$ \quad $e_{ij} = 1/n$ \quad $\alpha \approx .85$

$\tilde{P} = \alpha P + \alpha a v^T + (1 - \alpha)E$ (trivially irreducible)

- $\pi^T$ is now guaranteed to exist and be unique and power method is guaranteed to converge to $\pi^T$.

- Different $E = ev^T$ and $\alpha$ allow customization & speedup, yet rank-one update maintained; $\tilde{P} = \alpha P + (\alpha a + (1 - \alpha)e)v^T$

$\tilde{P} = \alpha P + (1 - \alpha)E = \begin{bmatrix}
1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\
1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\
1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\
1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\
1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \\
\end{bmatrix}$
Computing $\pi^T$

A Big Problem

Solve $\pi^T = \pi^T \tilde{P}$

$\pi^T(I - \tilde{P}) = 0$

(stationary distribution vector)

(too big for direct solves)
Google's PageRank is an eigenvector of a matrix of order 2.7 billion.

One of the reasons why Google is such an effective search engine is the PageRank algorithm, developed by Google's founders, Larry Page and Sergey Brin, when they were graduate students at Stanford University. PageRank is determined entirely by the link structure of the Web. It is recomputed about once a month and does not involve any of the actual content of Web pages or of any individual query. Then, for any particular query, Google finds the pages on the Web that match that query and lists those pages in the order of their PageRank.

Imagine surfing the Web, going from page to page by randomly choosing an outgoing link from one page to get to the next. This can lead to dead ends at pages with no outgoing links, or cycles around cliques of interconnected pages. So, a certain fraction of the time, simply choose a random page from anywhere on the Web. This theoretical random walk of the Web is a Markov chain or Markov process. The limiting probability that a dedicated random surfer visits any particular page is its PageRank. A page has high rank if it has links to and from other pages with high rank.

Let $W$ be the set of Web pages that can reached by following a chain of hyperlinks starting from a page at Google and let $n$ be the number of pages in $W$. The set $W$ actually varies with time, but in May 2002, $n$ was about 2.7 billion. Let $G$ be the $n$-by-$n$ connectivity matrix of $W$. The eigenvalues of $G$ are $0$ and $1$, and the largest eigenvalue is $1$.

It tells us that the largest eigenvalue of $A$ is equal to one and that the corresponding eigenvector, which satisfies the equation

$$x = Ax$$

exists and is unique to within a scaling factor. When this scaling factor is chosen so that

$$\sum_i x_i = 1$$

then $x$ is the state vector of the Markov chain. The elements of $x$ are Google's PageRank.

If the matrix were small enough to fit in MATLAB, one way to compute the eigenvector $x$ would be to start with a good approximate solution, such as the PageRanks from the previous month, and simply repeat the assignment statement

$$x = Ax$$

until successive vectors agree to within specified tolerance. This is known as the power method and is about the only possible approach for very large $n$. I'm not sure how Google actually computes PageRank, but one step of the power method would require one pass over a database of Web pages, updating weighted reference counts generated by the hyperlinks between pages.
Computing $\pi^T$

A Big Problem

Solve $\pi^T = \pi^T \bar{P}$ (stationary distribution vector)

$\pi^T(I - \bar{P}) = 0$ (too big for direct solves)

Start with $\pi^T_0 = e/n$ and iterate $\pi^T_{j+1} = \pi^T_j \bar{P}$ (power method)
Power Method to compute PageRank

\[ \pi_0^T = \frac{e^T}{n} \]

until convergence, do

\[ \pi_{j+1}^T = \pi_j^T \tilde{P} \]  
(dense computation)

end
Power Method to compute PageRank

\[ \pi^T_0 = \mathbf{e}^T / n \]

until convergence, do

\[ \times \quad \pi^T_{j+1} = \pi^T_j \tilde{P} \quad \text{(dense computation)} \]

\[ \bullet \quad \pi^T_{j+1} = \alpha \pi^T_j \tilde{P} + (1 - \alpha) \pi^T_j \mathbf{e} \mathbf{v}^T \quad \text{(sparser computation)} \]

end
Power Method to compute PageRank

\[ \pi_0^T = e^T/n \]

until convergence, do

\[ X \quad \pi_{j+1}^T = \pi_j^T \bar{P} \quad \text{(dense computation)} \]

\[ X \quad \pi_{j+1}^T = \alpha \pi_j^T \bar{P} + (1 - \alpha) \pi_j^T e v^T \quad \text{(sparser computation)} \]

\[ \bullet \quad \pi_{j+1}^T = \alpha \pi_j^T P + (\alpha \pi_j^T a + (1 - \alpha)) v^T \quad \text{(even less computation)} \]

end

\bullet \quad \text{P is very, very sparse with about 3-10 nonzeros per row.}

\bullet \quad \Rightarrow \text{one vector-matrix mult. is } O(nnz(P)) \approx O(n).
Convergence

Can prove $\lambda_2(\bar{P}) = \alpha$

($\Rightarrow$ asymptotic rate of convergence of PageRank method is rate at which $\alpha^k \rightarrow 0$)

Google

- uses $\alpha = 0.85$ (5/6, 1/6 interpretation)
- report 50-100 iterations til convergence
- still takes days to converge
PageRank Example

$$\pi^T = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ .03721 & .05396 & .04151 & .3751 & .206 & .2862 \end{pmatrix}$$

Global ranking of pages = [4 6 5 2 3 1]

Query-independent way of ranking relevant set
PageRank Issues

Spamming

- Link Farms
**What's News—World-Wide**

**Business and Finance**

NEWS CORP. and Liberty are no longer working together on a joint offer to take control of Hughes, with News Corp. proceeding on its own and Liberty considering an independent bid. The move threatens to cloud the process of finding a new owner for the GM unit. (Article on Page A3)

■ The SEC signaled it may file civil charges against Morgan Stanley, alleging it doled out IPO shares based partly on investors' commitments to buy more stock. (Article on Page C1)

■ Ahold's problems deepened as U.S. authorities opened inquiries into accounting at the Dutch company's U.S. Foodservice unit. (Articles on Page A2)

■ Consumer confidence fell to its lowest level since 1993, hurt by energy costs, the terrorism threat and a stagnant job market. (Article on Page A3)

■ The industrials rebounded on news of a possible settlement with

---

**Web Master**

As the Web spreads...

Total Internet users, by household, in millions

![Bar chart showing Internet users growth](chart)

<table>
<thead>
<tr>
<th>Year</th>
<th>Users, millions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td></td>
</tr>
</tbody>
</table>

**Google's U.S. presence expands**

Top search engines, in millions of unique visitors

<table>
<thead>
<tr>
<th>Engine</th>
<th>Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>39.4</td>
</tr>
<tr>
<td>Yahoo</td>
<td>38.6</td>
</tr>
<tr>
<td>MSN</td>
<td>36.8</td>
</tr>
<tr>
<td>AOL</td>
<td>22.0</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>13.3</td>
</tr>
<tr>
<td>Overture</td>
<td>6.4</td>
</tr>
</tbody>
</table>

**Top shopping-referral sites, in millions of referrals**

<table>
<thead>
<tr>
<th>Site</th>
<th>Referrals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dealline</td>
<td>2.50</td>
</tr>
<tr>
<td>BizRate</td>
<td>1.93</td>
</tr>
<tr>
<td>Overture</td>
<td>1.04</td>
</tr>
<tr>
<td>Epinions</td>
<td>0.78</td>
</tr>
<tr>
<td>CNET</td>
<td>0.76</td>
</tr>
</tbody>
</table>

---

**Bush to Seek up to $95 Billion To Cover Costs of War on Iraq**

By Greg Jaffe
And John D. McKinnon

WASHINGTON—The Bush administration is preparing supplemental spending requests totaling as much as $95 billion for a war with Iraq, its aftermath and new expenses to fight terrorism, officials said.

The total could be as low as $60 billion because Pentagon budget planners don't know how long a military conflict will last, whether U.S. allies will contribute more than token sums to the effort and what damage Saddam Hussein might do to his own country to retaliate against conquering forces.

Budget planners also are awaiting the outcome of an intense internal debate over whether to include $13 billion in the requests to Congress that the Pentagon says it needs to fund the broader war on terrorism, as well as for stepped up homeland security. The White House Office of Management and Budget argues that the money might not be necessary. President Bush, Defense Secretary Donald Rumsfeld and budget director Mitchell Daniels Jr. met yesterday to discuss the matter but didn't reach a final agreement. Mr. Rumsfeld plans to continue pressing his

---

**Cat and Mouse**

As Google Becomes Web's Gatekeeper, Sites Fight to Get In

Search Engine Punishes Firms That Try to Game System; Outlawing the 'Link Farms'

Exotic Leatherwear Gets Cut Off

By Michael Totty
And Meleyn Mangalindan

Joy Holman sells provocative clothing on the Web. She wants what nearly everyone doing business online wants: more exposure on Google. So from the time she launched exoticleatherwear.com last May, she tried all sorts of tricks to get her site to show up among the first listings when a user of Google Inc.'s popular search engine typed in "women's leatherwear" or "leather apparel." She buried hidden words in her Web pages intended to fool Google's computers. She signed up with a service that promised to have hundreds of sites link to her online store—thereby boosting a crucial measure in Google's system of ranking sites.

The techniques worked for a while.
Web Sites Fight for Prime Real Estate on Google

Continued From First Page

advertising that tried to capitalize on Google's formula for ranking sites. In fact, SearchKing was offering its clients a chance to boost their own Google rankings by buying keywords on more popular sites. SearchKing filed suit against the search company in federal court in Oklahoma, claiming that Google “purposefully defamed” SearchKing and its customers, damaging its reputation and hurting its advertising sales.

In court filings, the company said SearchKing “engaged in behavior that would lower the quality of Google search results” and alter the company’s ranking system.

Google, a closely held company founded by Stanford University graduate students Sergey Brin and Larry Page, says Web companies that want to rank high should concentrate on improving their Web pages rather than gaming its system. “When people try to game their own hands into their hands, that turns into a worse experience for users,” said Matt Cutts, a Google software engineer.

Coding Trickery

Efforts to outfox the search engines have been around since search engines first became popular in the early 1990s. Early tricks included inserting thousands of search terms in hidden coding, called “metatags.” The coding fools a search engine into identifying a site with popular words and phrases that may not actually appear on the site.

Another gimmick was hiding words or terms against a same-color background. The hiddenHTML tricks used by search engine spiders to find links to other sites, and thus their rankings. The hidden links were designed to be invisible to the human eye, but could be detected by search engines. They were particularly effective when the links were to sites that were popular with search engines, such as major news websites.

The more hidden links a site had, the higher it would rank in search engine results. But Google’s system, based on links, wasn’t fooled. In 2002, Google introduced a new algorithm called “Link Farming” that rewarded sites with high-quality links from other reputable sites, and penalized sites that engaged in tricks like hidden links.

On the surface, it seemed that the hidden links were a thing of the past. But in reality, the search engine tricks continued to evolve. In 2005, Google introduced a feature called “nofollow” that allowed site owners to tell search engines not to follow links on their pages. This helped to stem the tide of hidden links, but it was just a temporary solution.

In 2006, Google announced a new feature called “Google PR” that rewarded sites with high-quality links. This feature was designed to combat the use of hidden links, but it was not foolproof. Some site owners continued to use tricks to manipulate search engine rankings, and Google had to continually update its algorithms to combat these tactics.

In the end, the hidden links and other tricks used by search engine operators were a sign of the times. They were a testament to the lengths to which people would go to get ahead, and the need for search engines to continually evolve to stay ahead of the curve. But they were also a reminder of the importance of quality content, and the need for search engines to reward quality over tricks.
PageRank Issues

Spamming

- Link Farms
- Google Bombs
'Miserable failure' links to Bush

George W Bush has been Google bombed.

Web users entering the words "miserable failure" into the popular search engine are directed to the biography of the president on the White House website.

The trick is possible because Google searches more than just the contents of web pages - it also counts how often a site is linked to, and with what words.

Thus, members of an online community can affect the results of Google searches - called "Google bombing" - by linking their sites to a chosen one.

Weblogger Adam Mathes is credited with inventing the practice in 2001, when he used it to link the phrase "talentless hack" to a friend's website.

The search engine can be manipulated by a fairly small group of users, one report suggested.

Newsday newspaper says as few as 32 web pages with the words "miserable failure" link to the Bush biography.

The Bush administration has been on the receiving end of pointed Google bombs before.

In the run-up to the Iraq war, internet users manipulated Google so the phrase "weapons of mass destruction" led to a joke page saying "These Weapons of Mass Destruction cannot be displayed."

The site suggests "clicking the regime change button", or "If you are George Bush and typed the country's name in the address bar, make sure that it is spelled correctly (IRAQ)"

"Prank website

If you are George Bush and typed the country's name in the address bar, make sure that it is spelled correctly (IRAQ)"

E-mail services | Desktop ticker | Mobiles/PDAs |
10/27/2003 Archived Entry: "I'm taking part in a new web project..."

I'm taking part in a new web project...

From this day forth, I will refer to George W. Bush as a Miserable Failure at least once a day. Why, you ask? Well, someone came up with this great idea to link George W. Bush and Miserable Failure in popular search engines. If you have a blog or web site, help raise the link between George W. Bush and the phrase 'miserable failure' by copying this link and placing somewhere on your site or blog.

Thank you very much for your participation.

Replies: 16 people speak up

Great idea!

That is genius. I could add a few other keywords, like "pathetic". I will post it on my blog now...

Posted by Bjar @ 10/27/2003 10:06 PM NY

Miserable Failure? I'm down with that....

Stay tuned...

Posted by Drewcifer @ 10/28/2003 02:35 PM NY

Done!

Posted by Maru @ 10/28/2003 08:46 PM NY

That's great, another thing I think might be good to use: tax cuts for the wealthy....welfare for the wealthy, just my 2 cents.

Posted by dooda @ 10/29/2003 03:01 AM NY

Call me a liberal lemming, I guess. :) I'm in.

Posted by Bjar @ 10/29/2003 09:28 AM NY

The key is stating it in connection with terms that will be widely searched. It does no good to simply say "George Bush is a miserable failure" because no one will ever search for that. It might be fun at a parties to show how often the two are in the same sentence in a Google search, but otherwise it does little to advance the theme.

What will work is connecting it to frequent search terms, such as "Iraq policy". For instance "George Bush's Iraq Policy is a miserable failure."

The plan shouldn't be to link Miserable Failure to George Bush, but to link Miserable Failure to George Bush and two or three choice, frequently searched phrases.

Overture.com has a keyword suggestion tool that shows how many times certain terms are coming up in searches. Using that tool, I can determine that in September the search for "bush george iraq saddam" gets about 12 times more queries than "george bush iraq speech". "George bush iraq speech" gets a huge amount of hits compared to something like "george bush policy."

Since someone needs to write about three complete sentences using these terms based on verifiable search results and including the "miserable failure" phrase and then advocate for that exact usage.

According to Overture, the phrases "George Bush miserable failure" were not queried even once in their sample during the month just passed.

Posted by Joe Briefcase @ 10/29/2003 10:51 AM NY

...how about drunken, illiterate, mendacious, runt-like miserable failure?

Posted by tim @ 10/29/2003 11:58 AM NY

Hahaha, that's very productive. This is why everyone knows that liberals are stupid. They do stupid things.

Posted by Reek Stankleberry @ 10/29/2003 12:04 PM NY

...how about, instead of calling it lies--anyone can lie--how about calling it HORSEFEATHERS AND CODSWALLOP! Pin that on him too.

Replies: 16 people speak up
Google Search: miserable failure

Searched the web for miserable failure. Results 1 - 10 of about 257,000. Search took 0.08 seconds.
Tip: In most browsers you can just hit the return key instead of clicking on the search button.

Michael Moore.com
Wednesday, January 14th, 2004 I'll Be Voting For Wesley Clark /
Good-Bye Mr. Bush — by Michael Moore. Many of you have written ...
Description: Official site of the gadfly of corporations, creator of the film Roger and Me and the television show...
Category: Arts > Celebrities > M > Moore, Michael
www.michaelmoore.com/ - 43k - Cached - Similar pages

Biography of President George W. Bush
Home > President > Biography President George W. Bush En Español.
George W. Bush is the 43rd President of the United States. He ...
Description: Biography of the president from the official White House web site.
Category: Kids and Teens > School Time > ... > Bush, George Walker
www.whitehouse.gov/president/gwbbio.html - 29k - Cached - Similar pages

Biography of Jimmy Carter
Jimmy Carter aspired to make Government "competent and compassionate ...
Description: Short biography from the official White House site.
Category: Society > History > ... > Presidents > Carter, James Earl
www.whitehouse.gov/history/presidents/jc39.html - 36k - Cached - Similar pages

Senator Hillary Rodham Clinton: Online Office Welcome Page
Dear Friend,. Thank you for visiting my on-line office! I appreciate your interest in the issues before the United States Senate. ...
Description: Official US Senate web site of Senator Hillary Rodham Clinton (D - NY).
Category: Society > History > ... > First Ladies > Clinton, Hillary
clinton.senate.gov/ - 9k - Cached - Similar pages

BBC NEWS | Americas | 'Miserable failure' links to Bush
'Miserable failure' links to Bush. ... Prank website. Newsday newspaper says as few as 32 web pages with the words "miserable failure" link to the Bush biography. ...
news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - Cached - Similar pages

Atlantic Unbound | Politics & Prose | 2003.09.24
... Atlantic Unbound | September 24, 2003 Politics & Prose | by Jack Beatty
"A Miserable Failure" Will Bush be re-elected? Only if voters ...

miserable failure | Hillary Clinton | Hildebeest
... Miserable Failure. Quotes for the History Books. ... You may also want to check out the Miserable Failure Project. and the cuckolded dyke Project. and the ...
miserable-failure.blogspot.com/ - 60k - Cached - Similar pages

Dick Gephardt for President - Welcome
... to preserve some large part of the Bush tax cut. I think retaining
PageRank Issues

Spamming

- Link Farms
- Google Bombs

Updating

- The Google Dance
Below you will find the Google Dance results for the search keyword **pagerank**. If you notice that there are any differences in results between the different Google data centers then Google is in the middle of spidering the internet. It's that simple!

**www.google.com**

Enter your search:

**pagerank**

Google Search

I'm Feeling Lucky

1. Google Technology
2. Google Web Directory Help
3. Pagerank Explained, Google's PageRank Calculator, WebWorkshop
4. PageRank Explained Correctly with Examples
5. The Anatomy of a Search Engine
6. LinkAdage Auctions Link Exchange
7. PageRank is Dead (Jeremy Zawodny)
8. Google PageRank
9. The PageRank Citation Ranking:

Open Results in New Window

**www2.google.com**

Enter your search:

**pagerank**

Google Search

I'm Feeling Lucky

1. Google Technology
2. Google Web Directory Help
3. Pagerank Explained, Google's PageRank Calculator, WebWorkshop
4. PageRank Explained Correctly with Examples
5. The Anatomy of a Search Engine
6. LinkAdage Auctions Link Exchange
7. PageRank is Dead (Jeremy Zawodny)
8. Google PageRank

Open Results in New Window

**www3.google.com**

Enter your search:

**pagerank**

Google Search

I'm Feeling Lucky

1. Google Technology
2. Google Web Directory Help
3. Pagerank Explained, Google's PageRank Calculator, WebWorkshop
4. PageRank Explained Correctly with Examples
5. The Anatomy of a Search Engine
6. LinkAdage Auctions Link Exchange
7. PageRank is Dead (Jeremy Zawodny)
8. LinkAdage Auctions Link Exchange
9. Google PageRank

Open Results in New Window

Next Page>>

Google Dance Tool - Google Dance Results for pagerank
PageRank Issues

Spamming
- Link Farms
- Google Bombs

Updating
- The Google Dance

Speed Improvements
- Enhancing Power Method
Researchers Develop Techniques for Computing Google-Style Web Rankings Up to Five Times Faster

Speed-up may make "topic-sensitive" page rankings feasible

ARLINGTON, Va. — Computer science researchers at Stanford University have developed several new techniques that together may make it possible to calculate Web page rankings as used in the Google search engine up to five times faster. The speed-ups to Google's method may make it realistic to calculate page rankings personalized for an individual's interests or customized to a particular topic.

The Stanford team includes graduate students Sepandar Kamvar and Taher Haveliwala, noted numerical analyst Gene Golub and computer science professor Christopher Manning. They will present their first paper at the Twelfth Annual World Wide Web Conference (WWW2003) in Budapest, Hungary, May 20-24, 2003. The work was supported by the National Science Foundation, an independent federal agency that supports fundamental research and education in all fields of science and engineering.

Computing PageRank, the ranking algorithm behind the Google search engine, for a billion Web pages can take several days. Google currently ranks and searches 3 billion Web pages. Each personalized or topic-sensitive ranking would require a separate multi-day computation, but the payoff would be less time spent wading through irrelevant search results. For example, searching a sports-specific Google site for "Giants" would give more importance to pages about the New York or San Francisco Giants and less importance to pages about Jack and the Beanstalk.

"This work is a wonderful example of how NSF support for basic computer science research, including applied mathematics and algorithm research, has impacts in daily life," said NSF program officer Maria Zemankova. In the mid-1990s, an NSF digital library project and an NSF graduate fellowship also supported Stanford graduate students Larry Page and Sergey Brin while they developed what would become the Google search engine.

To speed up PageRank, the Stanford team developed a trio of techniques in numerical linear algebra. First, in the WWW2003 paper, they describe so-called "extrapolation" methods, which make some assumptions about the Web's link structure that aren't true, but permit a quick and easy computation of PageRank. Because the assumptions aren't true, the PageRank isn't exactly correct, but it's close and can be refined using the...
PageRank Issues

Spamming
- Link Farms
- Google Bombs

Updating
- The Google Dance

Speed Improvements
- Enhancing Power Method
- Personalized PageRank
A stealth start-up out of Stanford University is hoping to raise the heat on one of the toughest problems in Web search--and possibly out-Google Google in the process.

Kaltix was formed in recent months by three members of Stanford's PageRank team--a research group created to advance the mathematical algorithm developed by Google co-founder and Stanford alum Larry Page that cemented Google's fame.

PageRank has helped steer people to Web sites like no other search technology before it, harnessing the link structure of the Web to determine the most popular pages. Now, Kaltix hopes to improve upon PageRank, with an attempt to speed up the underlying PageRank computations.

That, in turn, could lay the groundwork for a breakthrough in a cutting-edge area of Web search development known as "personalization," which aims to sort search results based on the specific needs and interests of individuals, instead of the consensus approach pioneered by Google.

"Kaltix is a 'stealth mode' start-up... (leveraging) research done at Stanford University as well as several new technologies developed at Kaltix to provide large-scale personalized and context-sensitive search," a Kaltix representative said, declining to comment further.

Kaltix has disclosed few specifics about its plans or technology. But the company's general statements appear to place it in a sweet spot for innovation that's being pursued by all of the major search providers. Now that Web search has become a moneymaker for portals such as Yahoo and Microsoft's MSN, technologists from all the industry players are back in the labs developing formulas to personalize search.

Web companies outside the search industry have long made attempts to create personalization features, but most of these attempts have fallen short of expectations. Amazon.com, for example, regularly serves up book titles related to a visitor's previous purchases, which may no longer be relevant. A personalization feature offered through TiVo, a maker of video recording devices, was criticized when reports circulated that the device would recommend gay-themed television programs to viewers based on just a few program selections.

Despite these flawed attempts, developers continue to have faith that personalization technology can be created that will ultimately unleash marketing and revenue opportunities.

If search developers are successful in building such technology, they could help millions of people better...
Google Acquires Kaltix Corp.

New Technologies and Engineering Team Complement Google Search Engine

MOUNTAIN VIEW, Calif. - Sept. 30, 2003 - Google Inc. today announced it acquired Kaltix Corp., a Palo Alto, Calif.-based search technology start-up. Financial terms of the deal were not disclosed.

"Google and Kaltix share a common commitment to developing innovative search technologies that make finding information faster, easier and more relevant," said Larry Page, co-founder and president of Products at Google. "Kaltix is working on a number of compelling search technologies, and Google is the ideal vehicle for the continued development of these advancements."

Kaltix Corp. was formed in June 2003 and focuses on developing personalized and context-sensitive search technologies that make it faster and easier for people to find information on the web.

About Google
Google's innovative search technologies connect millions of people around the world with information every day. Founded in 1998 by Stanford Ph.D. students Larry Page and Sergey Brin, Google today is a top web property in all major global markets. Google's targeted advertising program, which is the largest and fastest growing in the industry, provides businesses of all sizes with measurable results, while enhancing the overall web experience for users. Google is headquartered in Silicon Valley with offices throughout North America, Europe, and Asia. For more information, visit www.google.com.

###

Google is a trademark of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

For further information:
Nathan Tyler
Google Inc.
+1 650-623-4311
nate@google.com
Conclusions

- Link Analysis has drastically improved web search!
- There are many exciting open problems for CSC and MATH majors to solve.
- Often the challenge lies not in the modeling or theory, but in the massive scale of the problem.
- The continual battle between search engines and search engine optimizers means that methods must constantly adapt and innovate.
- There is huge financial potential for industrious entrepreneurs!