Computing and Information Retrieval

The Big Picture

Amy N. Langville

Department of Mathematics
College of Charleston
Charleston, SC
Outline

• Introduction to Information Retrieval (IR)

• Traditional and Web Search

• Problems in Web Search

• Innovations
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
1872           Card catalog
1940s-1950s   Computer
1960s         Salton’s SMART system (trad. search)
1989          Berner-Lee’s WWW (web search)
Traditional Search

Two Primary Goals:

- Clustering documents
- Processing user queries
  - find similar documents
  - find similar terms
Vector Space Model (1960s and 1970s)

Gerard Salton’s Information Retrieval System
SMART: System for the Mechanical Analysis and Retrieval of Text
(Salton’s Magical Automatic Retriever of Text)

- turn \( n \) textual documents into \( n \) document vectors \( d_1, d_2, \ldots, d_n \)
- create term-by-document matrix \( A_{m \times n} = [d_1 | d_2 | \cdots | d_n] \)
- to retrieve info., create query vector \( q \), which is a pseudo-doc

**GOAL:** find doc. \( d_i \) closest to \( q \)

- **angular cosine** measure used: \( \delta_i = \cos \theta_i = q^T d_i / (\|q\|_2 \|d_i\|_2) \)
<table>
<thead>
<tr>
<th>Terms</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1: Baby's</td>
<td>D1: Infant &amp; Toddler First Aid</td>
</tr>
<tr>
<td>T2: Children's</td>
<td>D2: Babies &amp; Children's Room (For Your Home)</td>
</tr>
<tr>
<td>T3: Guide</td>
<td>D3: Child Safety at Home</td>
</tr>
<tr>
<td>T4: Health</td>
<td>D4: Your Baby's Health &amp; Safety: From Infant to Toddler</td>
</tr>
<tr>
<td>T5: Home</td>
<td>D5: Baby Proofing Basics</td>
</tr>
<tr>
<td>T6: Infant</td>
<td>D6: Your Guide to Easy Rust Proofing</td>
</tr>
<tr>
<td>T7: Proofing</td>
<td>D7: Beanie Babies Collector's Guide</td>
</tr>
<tr>
<td>T8: Safety</td>
<td></td>
</tr>
<tr>
<td>T9: Toddler</td>
<td></td>
</tr>
</tbody>
</table>
### Example from Berry’s book

<table>
<thead>
<tr>
<th>Terms</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1: Bab(y,ies,y’s)</td>
<td>D1: Infant &amp; Toddler First Aid</td>
</tr>
<tr>
<td>T2: Child(ren’s)</td>
<td>D2: Babies &amp; Children's Room (For Your Home)</td>
</tr>
<tr>
<td>T3: Guide</td>
<td>D3: Child Safety at Home</td>
</tr>
<tr>
<td>T4: Health</td>
<td>D4: Your Baby’s Health &amp; Safety : From Infant to Toddler</td>
</tr>
<tr>
<td>T5: Home</td>
<td>D5: Baby Proofing Basics</td>
</tr>
<tr>
<td>T6: Infant</td>
<td>D6: Your Guide to Easy Rust Proofing</td>
</tr>
<tr>
<td>T7: Proofing</td>
<td>D7: Beanie Babies Collector’s Guide</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
A &= \begin{pmatrix}
    d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \\
    t_1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\
    t_2 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
    t_3 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
    t_4 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
    t_5 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
    t_6 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
    t_7 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
    t_8 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
    t_9 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
\end{pmatrix}
\end{align*}

\[
q = \begin{pmatrix}
    1 \\
    0 \\
    0 \\
    1 \\
    0 \\
    0 \\
    0 \\
    0 \\
    0 \\
\end{pmatrix}, \quad
\delta = \begin{pmatrix}
    \delta_1 \\
    \delta_2 \\
    \delta_3 \\
    \delta_4 \\
    \delta_5 \\
    \delta_6 \\
    \delta_7 \\
\end{pmatrix}, \quad
\begin{pmatrix}
    0 \\
    .5774 \\
    0 \\
    .8944 \\
    .7071 \\
    0 \\
    .7071 \\
\end{pmatrix}
\]
Geometry of VSM for 3d vectors

vsmanimation.html
Latent Semantic Indexing (1990s)

Susan Dumais’s improvement to VSM = LSI

Idea: use low-rank approximation to $A$ to filter out noise

- Great Idea! 2 patents for Bell/Telcordia

Singular Value Decomposition

$A_{m \times n}$: rank $r$ term-by-document matrix

- SVD: $A = U \Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$

- LSI: use $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ in place of $A$

- Why?
  - reduce storage when $k << r$
  - filter out uncertainty, so that performance on text mining tasks (e.g., query processing and clustering) improves
LSI Demos

- Telcordia LSI Demo:  http://lsi.research.telcordia.com/lsi-bin/lsiQuery
- Netlib LSI Demo:  http://www.netlib.org/cgi-bin/lsiBook
Nonnegative Matrix Factorization

Idea: use low-rank approximation with nonnegative factors to improve LSI

\[ \mathbf{A}_k = \mathbf{U}_k \Sigma_k \mathbf{V}_k^T \]

\[ \mathbf{A}_k = \mathbf{W}_k \mathbf{H}_k \]

Daniel Lee and Sebastian Seung’s Nonnegative Matrix Factorization

(2000)
Properties of NMF

- can restrict $W, H$ to be sparse

- $W_k, H_k \geq 0 \Rightarrow$ immediate interpretation (additive parts-based rep.)

  EX: large $w_{ij}$'s $\Rightarrow$ basis vector $w_i$ is mostly about terms $j$

  EX: $h_{i1}$ how much $doc_1$ is pointing in the “direction” of topic vector $w_i$

$$A_k e_1 = W_k H e_1 = \begin{bmatrix} \vdots \\ w_1 \\ \vdots \end{bmatrix} h_{11} + \begin{bmatrix} \vdots \\ w_2 \\ \vdots \end{bmatrix} h_{21} + \cdots + \begin{bmatrix} \vdots \\ w_k \\ \vdots \end{bmatrix} h_{k1}$$
Interpretation of Basis Vectors

MED dataset ($k = 10$)

**Highest Weighted Terms in Basis Vector $W_1$:**
1. ventricle
2. aortic
3. septal
4. left
5. defect
6. regurgitation
7. ventricle
8. valve
9. cardiac
10. pressure

**Highest Weighted Terms in Basis Vector $W_2$:**
1. oxygen
2. flow
3. pressure
4. blood
5. cerebral
6. hypothermia
7. fluid
8. venous
9. arterial
10. perfusion

**Highest Weighted Terms in Basis Vector $W_5$:**
1. children
2. child
3. autistic
4. speech
5. group
6. early
7. visual
8. anxiety
9. emotional
10. autism

**Highest Weighted Terms in Basis Vector $W_6$:**
1. kidney
2. marrow
3. dna
4. cells
5. nephrectomy
6. unilateral
7. lymphocyte
8. bone
9. thymidine
10. rats
Interpretation of Basis Vectors

MED dataset \((k = 10)\)

\[
\text{doc}_5 \approx W_9 \begin{pmatrix}
\text{fatty} \\
\text{glucose} \\
\text{acids} \\
\text{ffa} \\
\text{insulin} \\
\vdots
\end{pmatrix} + 0.1646 + W_6 \begin{pmatrix}
\text{kidney} \\
\text{marrow} \\
\text{dna} \\
\text{cells} \\
\text{nephr.} \\
\vdots
\end{pmatrix} + 0.0103 + W_7 \begin{pmatrix}
\text{hormone} \\
\text{growth} \\
\text{hgh} \\
\text{pituitary} \\
\text{mg} \\
\vdots
\end{pmatrix} + 0.0045 + \cdots
\]
Computation of NMF

(Lee and Seung 2000)

Mean squared error objective function

\[
\min \| A - WH \|^2 \quad s.t. \quad W, H \geq 0
\]

\[
W = \text{abs} \left( \text{randn}(m,k) \right);
\]

\[
H = \text{abs} \left( \text{randn}(k,n) \right);
\]

for \( i = 1 : \text{maxiter} \)

\[
H = H .* \left( W^T A \right) ./ \left( W^T WH + 10^{-9} \right);
\]

\[
W = W .* \left( A H^T \right) ./ \left( WHH^T + 10^{-9} \right);
\]

end

Many parameters affect performance (k, obj. function, sparsity constraints, algorithm, etc.).

— NMF is not unique!
Looking for interesting work that matters to millions of people?

Google's mission:
Organize the world's information and make it universally accessible and useful.

To make this vision a reality, Google is looking for exceptional people who like to develop innovative new products, especially software engineers and tech-savvy product managers. Are you exceptional at what you do? Do you:

- Thrive on working in small teams to develop innovative products?
- Enjoy developing efficient new algorithms for processing tremendous amounts of data?
- Think it would be fun to write distributed systems that run on thousands of computers?
- Live to have the results of your work used and depended upon by millions of people every day?

If you're an outstanding software developer, computer scientist or product manager, read on and consider sending your resume and a brief cover letter to greatpeople@google.com.

If you know others who fit in this category, help us improve Google by forwarding this message and URL to them. The URL for this page is:

http://www.google.com/jobs/great-people-needed.html

What is it like to work at Google?

Working at Google means solving fascinating problems and making a positive difference in tens of millions of lives every day. This work has opened up interesting new areas for us and presented challenges that are not only new to us, but new to everyone in computing. These new problems require exceptional thinking and technical expertise to solve, but their solutions could dramatically improve the accessibility of information for everyone in the world. Here's a sampling of the kinds of things we work on at Google:
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
- Page Repository
- User
- Queries
- Results
- Indexing Module
- Indexes
- Content Index
- Structure Index
- Special-purpose indexes
- Query Module
- Ranking Module

- Elements
  - Indexing Module
  - Ranking Module
  - Query Module
  - Page Repository
  - User
  - WWW

- Processes
  - Crawler Module
  - Indexing Module
  - Query Module
  - Ranking Module

- Data
  - Indexes
  - Content Index
  - Structure Index
  - Special-purpose indexes
Indexing Wars

Actual Index King =
Internet Archive - http://web.archive.org
Search Stats—Google

- received over 0.5 billion searches per day in 2004
- stores an index of 8.1 billion webpages
- had over 60,000 servers in 2004
- estimated to use over 6,200 TB of disk space
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 architecture to compute the importance of each web page based on the votes cast by each page on the web. Each time a user enters a query, the result set is composed of the pages that PageRank identifies as the most important and relevant to that query.
The PageRank Idea

- Ranking is preassigned (An off-line calculation)
- 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

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

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

- $r_1(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_0(P)}{|P|}$

- $r_2(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_1(P)}{|P|}$

- \ldots

- $r_{j+1}(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_j(P)}{|P|}$
In Matrix Notation

After Step $j$

$$\pi^T_j = [r_j(P_1), r_j(P_2), \cdots, r_j(P_n)]$$

$$\pi^T_{j+1} = \pi^T_j H \quad \text{where} \quad h_{ij} = \begin{cases} 
1/|P_i| & \text{if } i \rightarrow j \\
0 & \text{o.w.}
\end{cases}$$
In Matrix Notation

After Step $j$

$$\pi^T_j = [r_j(P_1), r_j(P_2), \cdots, r_j(P_n)]$$

$$\pi^T_{j+1} = \pi^T_j H \quad \text{where} \quad h_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{o.w.} \end{cases}$$

$$H = \begin{pmatrix} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\ p_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ p_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ p_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ p_6 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$
In Matrix Notation

After Step \( j \)

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

\[
\pi_{j+1}^T = \pi_j^T H \quad \text{where} \quad h_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{o.w.} \end{cases}
\]

\[
H = \begin{pmatrix}
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}
\]

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

It’s Almost a Markov Chain

\( H \) has row sums = 1 for ND nodes, row sums = 0 for D nodes
In Matrix Notation

It’s Almost a Markov Chain

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

It’s Almost a Markov Chain

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

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

It’s Almost a Markov Chain

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

**Stochasticity Fix:** $S = H + av^T$.  
($a_i = 1$ for $i \in D$, 0, o.w.)

\[
S = \begin{bmatrix}
0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\
\frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\
\frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\
0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\
0 & 0 & 0 & \frac{1}{2} & 0 & \frac{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 = \frac{1}{6} \mathbf{e}^T
**In Matrix Notation**

**It’s Almost a Markov Chain**

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

Stochasticity Fix: $S = H + av^T$.  
($a_i = 1$ for $i \in D$, 0, o.w.)

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

- Each $\pi_j^T$ is a probability distribution vector  
($\sum_i r_j(P_i) = 1$)

- $\pi_{j+1}^T = \pi_j^T S$ 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

Long-run proportion of time on page $P_i$ is $\pi_i$ (Back button not a link)

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)  

Could get trapped into a cycle $(P_i \rightarrow P_j \rightarrow P_i)$  

(Back button not a link)  

($\pi^T$ not well defined)  

(No convergence)
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)  

Could get trapped into a cycle $(P_i \rightarrow P_j \rightarrow P_i)$  

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 $\begin{bmatrix} X & Y \\ 0 & Z \end{bmatrix}$. 
Random Surfer

Bored Surfer Enters Random URL

Irreducibility Fix: \( \mathbf{G} = \alpha \mathbf{S} + (1 - \alpha) \mathbf{E} \quad e_{ij} = 1/n \quad \alpha \approx 0.85 \)

\( \mathbf{G} = \alpha \mathbf{H} + \alpha \mathbf{a} \mathbf{v}^T + (1 - \alpha) \mathbf{E} \) (trivially irreducible)

- \( \pi^T \) 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: \[ G = \alpha S + (1 - \alpha)E \quad e_{ij} = 1/n \quad \alpha \approx 0.85 \]

\[ G = \alpha H + \alpha a v^T + (1 - \alpha)E \quad \text{(trivially irreducible)} \]

- \( \pi^T \) 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; \[ G = \alpha H + (\alpha a + (1 - \alpha)e)v^T \]

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

A Big Problem

Solve $\pi^T = \pi^T G$

$(\text{stationary distribution vector})$

$\pi^T (I - G) = 0$

$(\text{too big for direct solves})$
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 the graph. Let \( \|g_{ij}\| = 1 \) if page \( j \) links to page \( i \), and zero otherwise. It is a stochastic matrix with row sum one. Let \( x \) be the PageRank vector. For the stationary distribution, we have

\[
x = \lambda x,
\]

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 $\pi^T = \pi^T G$  

$\pi^T (I - G) = 0$  

Start with $\pi_0^T = e/n$ and iterate $\pi_{j+1}^T = \pi_j^T G$  

(stationary distribution vector)  

(too big for direct solves)  

(power method)
Power Method to compute PageRank

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

until convergence, do

\[ \pi_{j+1}^T = \pi_j^T G \]  

(dense computation)

end
Power Method to compute PageRank

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

until convergence, do

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

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

end
Power Method to compute PageRank

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

until convergence, do

- \( \pi_{j+1}^T = \pi_j^T G \) (dense computation)

- \( \pi_{j+1}^T = \alpha \pi_j^T S + (1 - \alpha) \pi_j^T e v^T \) (sparser computation)

- \( \pi_{j+1}^T = \alpha \pi_j^T H + (\alpha \pi_j^T a + (1 - \alpha)) v^T \) (even less computation)

end

- \( H \) is very, very sparse with about 3-10 nonzeros per row.

- \( \Rightarrow \) one vector-matrix mult. is \( O(nnz(H)) \approx O(n) \).
PageRank Example

\[ \pi^T = \begin{pmatrix} .03721 & .05396 & .04151 & .3751 & .206 & .2862 \end{pmatrix} \]

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

Query-independent way of ranking relevant set
Ranking by HITS

- 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)*
Sponsored Links

NCAA Bracket Contest: NCAA Bracket Contest at CollegeTournament.com
www.collegetournament.com

www.xposed.com

Results

Showing 1-10 of about 3,255,000:

NCAA: National Collegiate Athletic Association - Official Site
2004 NCAA Division I Men's Basketball Championship bracket announced...
www.ncaa.org/

Men's and Women's Basketball Polls
Division I Men's Basketball: The NCAA does not conduct a poll for Division I men's basketball and the NCAA's Division I Men's Basketball Committee...
www.ncaa.org/polls/m_w_basketball.html
[More Results from www.ncaa.org]

ESPN.com: Men's College Basketball
...to attend the EA SPORTS Maui Invitational Basketball Tournament, stay ... Wednesday ...
3:00 pm ... 1979 NCAA TOURNAMENT, MIDWEST REGIONAL 2ND...
sports.espn.go.com/ncaab/index

Men's Basketball - NCAA Sports.com
Live Game Video NCAA March Madness on Demand brings you LIVE video of the Men's Basketball tournament. Division I...
www.ncaasports.com/basketball/mens

NCAA Basketball
Live Game Video NCAA March Madness on Demand brings you LIVE video of the Men's Basketball tournament. Division I Men's Basketball...
www.ncaasports.com/
[More Results from www.ncaasports.com]

D3hoops.com: The definitive resource for Division III men's and...
The definitive resource for Division III men's and women's basketball ... previews: M | W Final Four: M | W Stats (NCAA site): M | W NCAA rankings: M...
www.d3hoops.com/

Women's Basketball Coaches Association
March 14 Selection Sunday for the NCAA Division I Women's Basketball Tournament March 16 NAIA DII Women's Championship March 19 NCAA DIII...
www.wbcaa.org/
[More Results from www.wbcaa.org]

CollegeRPI.com - College Basketball Rating Percentage Index (RPI)
The most accurate independent duplication of the NCAA's Rating Percentage Index...
www.collegerpi.com/

College Basketball by CollegeHoopsnet.com
Player of the Week. NCAA Tournament. Conference Tourneys. Basketball Tickets. Recruiting Coverage. Basketball Store. NBA Draft...
www.collegehoopsnet.com/

CBS.SportsLine.com - NCAA Basketball Home
College Basketball coverage including NCAA news, scores, standings, stats, schedules, injuries, polls, team and player news, NCAA basketball...
www.sportsline.com/collegebasketball/

Results Pages: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 >

© 2004 Ask Jeeves, Inc.
Teoma is a service mark of Ask Jeeves, Inc. All rights reserved.
Terms of Service
HITS Algorithm
Hypertext Induced Topic Search  
(J. Kleinberg 1998)

Determine Authority & Hub Scores

- $a_i = \text{authority score for } P_i$
- $h_i = \text{hub score for } P_i$

Successive Refinement

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

- $A = L^T L$ \quad $a_k = Aa_{k-1} \rightarrow \text{e-vector}$
- $H = LL^T$ \quad $h_k = Hh_{k-1} \rightarrow \text{e-vector}$
1. Find relevant set by consulting inverted file
2. Build neighborhood graph
3. Compute authority & hub scores for just the neighborhood
1. Relevant set = [1, 6]

2. Neighborhood graph $N$

3. Compute authority & hub scores.

Adjacency matrix for $N = L =$

$$
\begin{pmatrix}
1 & 2 & 3 & 5 & 6 & 10 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 \\
2 & 1 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 0 & 0 & 0 & 1 & 0 \\
5 & 0 & 0 & 0 & 0 & 0 & 0 \\
6 & 0 & 0 & 1 & 1 & 0 & 0 \\
10 & 0 & 0 & 0 & 0 & 1 & 0 \\
\end{pmatrix}
$$
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 \( \mathbf{h} \)

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

Other Rankings

- Consensus Ranking
Web Search Problems

Spam

- Link Farms
**What's News—**

**Business and Finance World-Wide**

**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.** Fleming said the SEC upgraded to a formal investigation an inquiry into the food wholesaler's trade practices with suppliers. (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 hopes of a robust 2003.** (Article on Page A3)

---

**Web Master**

As the Web spreads...

Total Internet users, by household, in millions

![Graph showing Internet user growth from 1997 to 2003](image)

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

Top search engines, in millions of unique visitors

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>39.4</td>
</tr>
<tr>
<td>Yahoo Search</td>
<td>38.6</td>
</tr>
<tr>
<td>MSN Search</td>
<td>36.8</td>
</tr>
<tr>
<td>AOL Search</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>Google</td>
<td>12.61</td>
</tr>
<tr>
<td>DeallTime</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'

Exoticleatherwear Gets Cut Off

**By MICHAEL TOTTY**

And MYLENE 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 Page 1

advertising that tried to capitalize on Google's formula for ranking sites. In effect, SearchKing was offering its clients a chance to boost their own Google rankings by buying links on more-popular sites. SearchKing sued against the search company in federal court in Oklahoma, claiming that Google "purposefully devalued" 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 funded 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 take scoring into their own hands, that turns into a worse experience for users," says Matt Cutts, a Google software engineer.

Citing Trickery

Efforts to outfox the search engines have been around since search engines first became popular in the early 1990s. Early tricks included hiding 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 hidden words were designed to fool search engines that relied heavily on the number of times a word or phrase appeared in ranking a site. But Google's system, based on links, wasn't fooled.

Mr. Brin, 29, one of Google's two founders and now its president of technology, opened a pop-up shop to the Exotic Leather Wear store in Mesa, Ariz., where he quickly learned the importance of appearing near the top of search-engine results, especially on Google. She boned up on search techniques, including online discussion groups dedicated to search engines and reading what material she could find on the Web.

At first, Ms. Holman limited herself to modest changes, such as loading her page with hidden metatag coding that would help steer a search toward her site when a user entered words such as "halter tops" or "leather pants. "Since Google doesn't give much weight to metatags in determining its rankings, the efforts had little effect on her search results.

She then received an e-mail advertisement from AutomatedLinks.com, a Wirral, England, company that promised to send traffic "through the roof" by linking more than 2,000 Web sites to hers. Aside from attracting customers, the links were designed to improve her rankings on search engines.

So she signed up AutomatedLinks and placed an advertisement within its network in July. She says she immediately started seeing traffic on her site, which was steadily increasing because links to her site remained on the sites of other AutomatedLinks customers. She paid the company fees, which were charged on a per-click basis of 50 cents.

In theory, when Google encounters the AutomatedLinks code, it treats it as a legitimate referral to other sites and counts it in topting the sites' popularity. But users who signed up with AutomatedLinks in J uly, she says, on an online discussion group to which Google objected to such link arrangements. She says she immediately stripped the code from her Web pages.

For a while her site gradually worked its way up in Google search rankings, but it has steadily improved because links to her site still remained on the sites of other AutomatedLinks customers. Sometimes in November, her site was suddenly no longer appearing among the top results. Her orders plunged as much as 80%.

Ms. Holman, who had maintained AutomatedLinks, says she has been unable to get answers. But in the last few months, other AutomatedLinks customers say they have seen their sites apparently penalized by Google. Graham McLeay, who runs a small chauffeur service, said his site rose to the first page of Google's rankings in mid-March, but fell back to half during the two months he believes his site was penalized by Google.

The high-stakes battle between Google and the optimizers can leave some Web site owners confused. "I don't know how people are supposed to judge what is right and what is wrong," she says. AutomatedLinks didn't respond to requests for comment. Google declined to comment on the case. But Mr. Cutts, the Google engineer, warns that the rules are clear and that it's better to follow them rather than try to get a problem fixed after a site has been penalized. "We want to return the most relevant pages we can," Mr. Cutts says. "The best way for a site owner to do that is follow our guidelines."

Crackdown

Google has been stepping up its enforcement since 2001. It warned Webmasters that using trickery could get their sites kicked out of the Google index and it provided a list of forbidden activities, including hiding text and "link schemes," such as the link farms. Google also warned against "cloaking"—hiding a search engine a page that's designed to score well while giving visitors a different, more attractive page—or creating multiple Web addresses that take visitors to a single site.

Fiat Patriarch

Is Set to Become

By Alessandra Galloni

ROME—Umberto Agnelli is due to name Fiat SpA chairman on Friday, a move that will allow the family's bloated conglomerate to make serious inroads on an 11th-hour resuscitation of its unprofitable car unit.

Mr. Agnelli, the 68-year-old brother of Fiat patriarch Gianni Agnelli, who last month was widely expected to step over from current chairman, Mr. Agnelli, who has served as chairman since 1993, has long been seen as a man who can drive change.
Web Search Problems

Spam

- 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)."

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
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...

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

Stay tuned...

Done!

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

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

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 biography" gets a huge amount of hits compared to something like "george bush policy".

So 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.

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

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

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

Web Images Groups Directory News
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
Web Search Problems

Spam

- Link Farms
- Google Bombs
  - search for algorithmic solutions that scale up

Social: getting surfers to use relevance feedback

- Specialized Tools: Froogle, Scholar
Web Search Problems

Spam

- Link Farms
- Google Bombs

Social: getting surfers to use relevance feedback

- Specialized Tools: Froogle, Scholar
- Personalization: www.a9.com
Web Search Problems

Spam

- Link Farms
- Google Bombs

Social: getting surfers to use relevance feedback

- Specialized Tools: Froogle, Scholar
- Personalization: www.a9.com

No human interaction, no librarian for mind-reading
Bad Results because ...

User’s Fault

- poor query
- typo
Bad Results because ... 

**User’s Fault**
- poor query
- typo

**Engine’s Fault**
- spam
- small index
Bad Results because ...

**User’s Fault**
- poor query
- typo

**Engine’s Fault**
- spam
- small index

**Web Community’s Fault**
- no quality pages posted on user’s query ⇒ do your part
Innovation

- metatag for Library of Congress #  (“nothing like a good book”)
Innovation

- metatag for Library of Congress #  (“nothing like a good book”)
Innovation

- metatag for Library of Congress # ("nothing like a good book")
- phonetic search in audio collections
Innovation

- metatag for Library of Congress # ("nothing like a good book")
- phonetic search in audio collections
- relevance feedback
  - edit distance for typos
  - synonyms (find similar terms using VSM)
Innovation

- metatag for Library of Congress # ("nothing like a good book")
- phonetic search in audio collections
- relevance feedback
  - edit distance for typos
  - synonyms (find similar terms using VSM)
- image search: google
Conclusions

- Link Analysis has drastically improved web search!
- There are many exciting open problems for computational scientists 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!