Outline

• Introduction to Information Retrieval
• Elements of a Search Engine
• Link Analysis
• Current Issues in Web Search
• PageRank and You
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
1940s-1950s Card catalog
1960s Salton’s SMART system
1989 Berner-Lee’s WWW
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.
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
- Query Module
- Ranking Module
- Indexing Module
- Indexes
- Special-purpose indexes
- Structure Index
- Content Index
- Query-independent
- Results
Submitting your Site to a Search Engine

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

Diagram:
- WWW (web) connects to Crawler Module
- Crawler Module connects to Page Repository
- Page Repository connects to Indexing Module
- Indexing Module connects to Ranking Module
- Ranking Module connects to Query Module
- User submits Queries to Query Module
- Query Module provides Results to User
Share your place on the net with us.

We add and update new sites to our index each time we crawl the web, and we invite you to submit your URL here. We do not add all submitted URLs to our index, and we cannot make any predictions or guarantees about when or if they will appear.

Please enter your full URL, including the http:// prefix. For example: http://www.google.com/. You may also add comments or keywords that describe the content of your page. These are used only for our information and do not affect how your page is indexed or used by Google.

Please note: Only the top-level page from a host is necessary; you do not need to submit each individual page. Our crawler, Googlebot, will be able to find the rest. Google updates its index on a regular basis, so updated or outdated link submissions are not necessary. Dead links will 'fade out' of our index on our next crawl when we update our entire index.

URL: 
Comments: 
Optional: To help us distinguish between sites submitted by individuals and those automatically entered by software robots, please type the squiggly letters shown here into the box below.

Add URL

Need to remove a site from Google? For more information, click here.
Elements of a Web Search Engine
Indexing Wars

Actual Index King =
Internet Archive - http://web.archive.org
Elements of a Web Search Engine

WWW

Crawler Module

Page Repository

User

Indexing Module

Queries

Results

Query Module

Indexes

Ranking Module

Content Index

Structure Index

Special-purpose indexes

query-independent
Ranking on the Web

the pre-1998 Web

- border patrol: 4; 567; 809; 1103;
- hezbollah: 9; 12; 339; 942; 15158;
- global warming: 178; 12980; 445532;
Index

$k$-step transition matrix, 179

a vector, 37, 38, 75, 80
A9, 142
absolute error, 104
absorbing Markov chains, 185
absorbing states, 185
accuracy, 79–80
adaptive PageRank method, 89–90
Adar, Eytan, 146
adjacency list, 77
adjacency matrix, 33, 76, 116, 132, 169
advertising, 45
aggregated chain, 197
aggregated chains, 195
aggregated transition matrix, 105
aggregated transition probability, 197
aggregation, 94–97
approximate, 102–104
exact, 104–105
exact vs. approximate, 105–107
iterative, 107–109
partition, 109–112
aggregation in Markov chains, 197
aggregation theorem, 105
Aitken extrapolation, 91
Alexa traffic ranking, 138
algebraic multiplicity, 157
algorithm
PageRank, 40
Aitken extrapolation, 92
dangling node PageRank, 82, 83
HITS, 116
iterative aggregation updating, 108
personalized PageRank power method, 49
quadratic extrapolation, 93
query-independent HITS, 124
$\alpha$ parameter, 37, 38, 41, 47–48
Amazon’s traffic rank, 142
anchor text, 48, 54, 201
Ando, Albert, 110
aperiodic, 36, 133
aperiodic Markov chain, 176
Application Programming Interface (API), 65, 73, 97
approximate aggregation, 102–104
arc, 201
Arrow, Kenneth, 136
asymptotic rate of convergence, 41, 47, 101, 119, 125
Atlas of Cyberspace, 27
authority, 29, 201
authority Markov chain, 132
authority matrix, 117, 201
authority score, 115, 201
authority vector, 201
Babbage, Charles, 75
back button, 84–86
BadRank, 141
Barabasi, Albert-Laszlo, 30
Berry, Michael, 7
bibliometrics, 32, 123
bipartite undirected graph, 131
BlockRank, 94–97, 102
blog, 55, 144–146, 201
Boldi, Paolo, 79
Boolean model, 5–6, 201
bounce back, 84–86
bowtie structure, 134
Brezinski, Claude, 92
Brin, Sergey, 25, 205
Brown, Murray, 7
Bush, Vannevar, 3, 10
Campbell, Lord John, 23
canonical form, reducible matrix, 182
censored chain, 104
censored chains, 194
censored distribution, 104, 195
censored Markov chain, 194
censorship, 146–147
Cesàro sequence, 162
Cesàro summability, stochastic matrix, 182
characteristic polynomial, 120, 156
Chebyshev extrapolation, 92
Chien, Steve, 102
cloaking, 44
clustering search results, 142–143
citation, 123, 201
citation, 123, 201
Collatz–Wielandt formula, 168, 172
core, 30
collapsed matrix storage, 76
collection, 59, 71, 155
Condorcet, 136
coupled Markov chains, 164
Ranking on the Web

the pre-1998 Web

- border patrol: 4; 567; 809; 1103; . . . (8,700,000 in total)
- hezbollah: 9; 12; 339; 942; 15158; . . . (15,100,000 in total)
- global warming: 178; 12980; 445532; . . . (33,200,000 in total)
Ranking on the Web

the pre-1998 Web

• border patrol: 4; 567; 809; 1103; . . . (8,700,000 in total)
• hezbollah: 9; 12; 339; 942; 15158; . . . (15,100,000 in total)
• global warming: 178; 12980; 445532; . . . (33,200,000 in total)

too many results per search term
easily spammed
1998: enter Link Analysis

- uses hyperlink structure to focus the relevant set
1998: enter Link Analysis

- uses hyperlink structure to focus the relevant set
- combine IR score with popularity score
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and \textit{quality} of votes cast for that page.

Hyperlink as vote

Markov chain
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote
Ranking with a Random Surfer

• Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and quality of votes cast for that page.

Hyperlink as vote
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and quality of votes cast for that page.

Hyperlink as vote
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and \textit{quality} of votes cast for that page.

Hyperlink as vote
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and quality of votes cast for that page.

Hyperlink as vote
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and quality of votes cast for that page.

Hyperlink as vote

page 2 is a dangling node
Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and \textit{quality} of votes cast for that page.

\textbf{Hyperlink as vote}

surfer “teleports”
Ranking with a Random Surfer

- If a page is “important,” it gets lots of votes from other important pages, which means the random surfer visits it often.

- Simply count the number of times, or proportion of time, the surfer spends on each page to create ranking of webpages.
Ranking with a Random Surfer

- If a page is “important,” it gets lots of votes from other important pages, which means the random surfer visits it often.

- Simply count the number of times, or proportion of time, the surfer spends on each page to create ranking of webpages.

<table>
<thead>
<tr>
<th>Proportion of Time</th>
<th>Ranked List of Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 1 = .04</td>
<td>Page 4</td>
</tr>
<tr>
<td>Page 2 = .05</td>
<td>Page 6</td>
</tr>
<tr>
<td>Page 3 = .04</td>
<td>Page 5</td>
</tr>
<tr>
<td>Page 4 = .38</td>
<td>Page 2</td>
</tr>
<tr>
<td>Page 5 = .20</td>
<td>Page 1</td>
</tr>
<tr>
<td>Page 6 = .29</td>
<td>Page 3</td>
</tr>
</tbody>
</table>
Ranking with a Random Surfer

- If a page is “important,” it gets lots of votes from other important pages, which means the random surfer visits it often.

- Simply count the number of times, or proportion of time, the surfer spends on each page to create ranking of webpages.

**Proportion of Time**

<table>
<thead>
<tr>
<th>Page</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 1</td>
<td>0.04</td>
</tr>
<tr>
<td>Page 2</td>
<td>0.05</td>
</tr>
<tr>
<td>Page 3</td>
<td>0.04</td>
</tr>
<tr>
<td>Page 4</td>
<td>0.38</td>
</tr>
<tr>
<td>Page 5</td>
<td>0.20</td>
</tr>
<tr>
<td>Page 6</td>
<td>0.29</td>
</tr>
</tbody>
</table>

**Ranked List of Pages**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Page 4</td>
</tr>
<tr>
<td>2</td>
<td>Page 6</td>
</tr>
<tr>
<td>3</td>
<td>Page 5</td>
</tr>
<tr>
<td>4</td>
<td>Page 2</td>
</tr>
<tr>
<td>5</td>
<td>Page 1</td>
</tr>
<tr>
<td>6</td>
<td>Page 3</td>
</tr>
</tbody>
</table>

query-independent
Web Graphs

CSC and MATH problems here:

- store adjacency matrix
- update adjacency matrix
- visualize web graph
- locate clusters in graph
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 underlying link structure as its foundation. It assigns each web page a numerical weight or value, called PageRank, that reflects the importance of the page in the overall structure of the web.
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 $W$ with $G_{ij}$ the fraction of the time a surfer who is on page $i$ clicks on a link to page $j$. Let $x$ be the stationary state or limiting probability vector.

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.
Search Issues

Spamming

- Link Farms
**What’s News—**

**Business and Finance**

**World-Wide**

- **Bush is preparing** to present Congress a huge bill for Iraq costs.
  - The total could run to $95 billion depending on the length of the possible war and occupation. As horsetrading began at the U.N. to win support for a war resolution, the president again made clear he intends to act with or without the world body's imprimatur. Arms inspectors said Baghdad provided new data, including a report of a possible biological bomb. Gen. Franks assumed command of the war-operations center in Qatar. Allied warplanes are aggressively taking out missile sites that could threaten the allied troop buildup. (Column 4 and Pages A4 and A6)

- **Turkey’s parliament debated legislation** to let the U.S. deploy 62,000 to open a northern front. Kurdish soldiers lined roads in a show of force as U.S. officials traveled into Iraq's north for an opposition conference.

- **Powell said North Korea hasn’t restarted a reactor and plutonium-processing facility** at Yongbyon, hinting such forbearance might constitute an overture. But saber rattling continued a day after a missile test timed for the inauguration in Seoul. Pyongyang accused U.S. spy planes of violating its airspace and told its army to prepare for U.S. attack. (Page A11)

- **The FBI came under withering bipartisan criticism in a Senate Judiciary report** in which Sen. Specter

---

**Web Master**

**As the Web spreads...**

<table>
<thead>
<tr>
<th>Total Internet users, by household, in millions</th>
<th>Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>50.5</td>
</tr>
<tr>
<td>2002</td>
<td>60.0</td>
</tr>
<tr>
<td>2003</td>
<td>70.0</td>
</tr>
<tr>
<td>2004</td>
<td>80.0</td>
</tr>
<tr>
<td>2005</td>
<td>90.0</td>
</tr>
<tr>
<td>2006</td>
<td>100.0</td>
</tr>
<tr>
<td>2007</td>
<td>110.0</td>
</tr>
<tr>
<td>2008</td>
<td>120.0</td>
</tr>
<tr>
<td>2009</td>
<td>130.0</td>
</tr>
<tr>
<td>2010</td>
<td>140.0</td>
</tr>
<tr>
<td>2011</td>
<td>150.0</td>
</tr>
<tr>
<td>2012</td>
<td>160.0</td>
</tr>
<tr>
<td>2013</td>
<td>170.0</td>
</tr>
<tr>
<td>2014</td>
<td>180.0</td>
</tr>
<tr>
<td>2015</td>
<td>190.0</td>
</tr>
<tr>
<td>2016</td>
<td>200.0</td>
</tr>
<tr>
<td>2017</td>
<td>210.0</td>
</tr>
<tr>
<td>2018</td>
<td>220.0</td>
</tr>
<tr>
<td>2019</td>
<td>230.0</td>
</tr>
<tr>
<td>2020</td>
<td>240.0</td>
</tr>
<tr>
<td>2021</td>
<td>250.0</td>
</tr>
<tr>
<td>2022</td>
<td>260.0</td>
</tr>
<tr>
<td>2023</td>
<td>270.0</td>
</tr>
<tr>
<td>2024</td>
<td>280.0</td>
</tr>
<tr>
<td>2025</td>
<td>290.0</td>
</tr>
<tr>
<td>2026</td>
<td>300.0</td>
</tr>
<tr>
<td>2027</td>
<td>310.0</td>
</tr>
<tr>
<td>2028</td>
<td>320.0</td>
</tr>
<tr>
<td>2029</td>
<td>330.0</td>
</tr>
<tr>
<td>2030</td>
<td>340.0</td>
</tr>
<tr>
<td>2031</td>
<td>350.0</td>
</tr>
<tr>
<td>2032</td>
<td>360.0</td>
</tr>
<tr>
<td>2033</td>
<td>370.0</td>
</tr>
<tr>
<td>2034</td>
<td>380.0</td>
</tr>
<tr>
<td>2035</td>
<td>390.0</td>
</tr>
<tr>
<td>2036</td>
<td>400.0</td>
</tr>
<tr>
<td>2037</td>
<td>410.0</td>
</tr>
<tr>
<td>2038</td>
<td>420.0</td>
</tr>
<tr>
<td>2039</td>
<td>430.0</td>
</tr>
<tr>
<td>2040</td>
<td>440.0</td>
</tr>
<tr>
<td>2041</td>
<td>450.0</td>
</tr>
<tr>
<td>2042</td>
<td>460.0</td>
</tr>
<tr>
<td>2043</td>
<td>470.0</td>
</tr>
<tr>
<td>2044</td>
<td>480.0</td>
</tr>
<tr>
<td>2045</td>
<td>490.0</td>
</tr>
<tr>
<td>2046</td>
<td>500.0</td>
</tr>
<tr>
<td>2047</td>
<td>510.0</td>
</tr>
<tr>
<td>2048</td>
<td>520.0</td>
</tr>
<tr>
<td>2049</td>
<td>530.0</td>
</tr>
<tr>
<td>2050</td>
<td>540.0</td>
</tr>
<tr>
<td>2051</td>
<td>550.0</td>
</tr>
<tr>
<td>2052</td>
<td>560.0</td>
</tr>
<tr>
<td>2053</td>
<td>570.0</td>
</tr>
<tr>
<td>2054</td>
<td>580.0</td>
</tr>
<tr>
<td>2055</td>
<td>590.0</td>
</tr>
<tr>
<td>2056</td>
<td>600.0</td>
</tr>
<tr>
<td>2057</td>
<td>610.0</td>
</tr>
<tr>
<td>2058</td>
<td>620.0</td>
</tr>
<tr>
<td>2059</td>
<td>630.0</td>
</tr>
<tr>
<td>2060</td>
<td>640.0</td>
</tr>
<tr>
<td>2061</td>
<td>650.0</td>
</tr>
<tr>
<td>2062</td>
<td>660.0</td>
</tr>
<tr>
<td>2063</td>
<td>670.0</td>
</tr>
<tr>
<td>2064</td>
<td>680.0</td>
</tr>
<tr>
<td>2065</td>
<td>690.0</td>
</tr>
<tr>
<td>2066</td>
<td>700.0</td>
</tr>
<tr>
<td>2067</td>
<td>710.0</td>
</tr>
<tr>
<td>2068</td>
<td>720.0</td>
</tr>
<tr>
<td>2069</td>
<td>730.0</td>
</tr>
<tr>
<td>2070</td>
<td>740.0</td>
</tr>
<tr>
<td>2071</td>
<td>750.0</td>
</tr>
<tr>
<td>2072</td>
<td>760.0</td>
</tr>
<tr>
<td>2073</td>
<td>770.0</td>
</tr>
<tr>
<td>2074</td>
<td>780.0</td>
</tr>
<tr>
<td>2075</td>
<td>790.0</td>
</tr>
<tr>
<td>2076</td>
<td>800.0</td>
</tr>
<tr>
<td>2077</td>
<td>810.0</td>
</tr>
<tr>
<td>2078</td>
<td>820.0</td>
</tr>
<tr>
<td>2079</td>
<td>830.0</td>
</tr>
<tr>
<td>2080</td>
<td>840.0</td>
</tr>
<tr>
<td>2081</td>
<td>850.0</td>
</tr>
<tr>
<td>2082</td>
<td>860.0</td>
</tr>
<tr>
<td>2083</td>
<td>870.0</td>
</tr>
<tr>
<td>2084</td>
<td>880.0</td>
</tr>
<tr>
<td>2085</td>
<td>890.0</td>
</tr>
<tr>
<td>2086</td>
<td>900.0</td>
</tr>
<tr>
<td>2087</td>
<td>910.0</td>
</tr>
<tr>
<td>2088</td>
<td>920.0</td>
</tr>
<tr>
<td>2089</td>
<td>930.0</td>
</tr>
<tr>
<td>2090</td>
<td>940.0</td>
</tr>
<tr>
<td>2091</td>
<td>950.0</td>
</tr>
<tr>
<td>2092</td>
<td>960.0</td>
</tr>
<tr>
<td>2093</td>
<td>970.0</td>
</tr>
<tr>
<td>2094</td>
<td>980.0</td>
</tr>
<tr>
<td>2095</td>
<td>990.0</td>
</tr>
<tr>
<td>2096</td>
<td>1000.0</td>
</tr>
</tbody>
</table>

**Google’s U.S. presence expands**

<table>
<thead>
<tr>
<th>Top search engines, in millions of unique visitors</th>
<th>Estimates</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>38.6</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>Google</th>
<th>12.61</th>
</tr>
</thead>
<tbody>
<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 case.

The techniques worked. For up—

---

**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 leather 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 effect, Searchking was offering its clients a chance to boost their own Google rankings by buying AdWords on more popular sites. SearchKing filed suit 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 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, that turns into a worse experience for users," says 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 stuffing thousands of search terms in hidden code, called "metatags." The coding fooled a search engine into identifying visitors 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 text devised search engine 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 new ventures, apologized to a San Francisco search-engine conference in 2000 that Google wasn't worried about having its results clogged with irrelevant results because its search methods couldn't be manipulated.

That didn't stop search optimizers from finding other ways to outfox the system. Attempts to manipulate Google's results even became a sport, called Googling. More than 10,000 Web sites have employed one or two tricks.
Link Farms

SEO client

Lincoln

Jefferson

Truman

Reagan

Clinton

Kennedy

Bush

Washington
Search 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

SEEN ALSO:
WMD spoof is internet hit
04 Jul 03  |  West Midlands
Google hit by link bombers
13 Mar 02  |  Science/Nature

TOP AMERICAS STORIES NOW
US army battles to keep soldiers
Report backs US Catholic bishops
Envoys bid to ease BSE fears
Protests widen over sky marshals

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...

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

Stay tuned...

Done!

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.

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 amounts 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

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
Google Bomb

Bob's page

miserable failure

Jim's blog

miserable failure

G. W. Bush Bio webpage

Kim's blog

miserable failure
Search Issues

Spamming

• Link Farms
• Google Bombs

Personalization

• Google’s psearch, A9, Kartoo
Search Issues

Spamming
- Link Farms
- Google Bombs

Personalization
- Google’s psearch, A9, Kartoo

Privacy
- Great Firewall of China
- AOL Data Leak
A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

AOL removed the search data from its site over the weekend and apologized for its release, saying it was an unauthorized move by a team that had hoped it would benefit academic researchers.

But the detailed records of searches conducted by Ms. Arnold and 657,000 other Americans, copies of which continue to circulate online, underscore how much people unintentionally reveal about themselves when they use search engines — and how risky it can be for companies like AOL, Google and Yahoo to compile such data.

Those risks have long pitted privacy advocates against online marketers and other Internet companies seeking to profit from the Internet's unique ability to track the comings and goings of users, allowing for more focused and therefore more lucrative advertising.

But the unintended consequences of all that data being compiled, stored and cross-linked are what Marc Rotenberg, the executive director of the Electronic Privacy Information Center, a privacy rights group in Washington, called "a ticking privacy time bomb."

Mr. Rotenberg pointed to Google's own joust earlier this year with the Justice...
Search Issues

Spamming
- Link Farms
- Google Bombs

Personalization
- Google’s psearch, A9, Kartoo

Privacy
- Great Firewall of China
- AOL Data Leak

Data Fusion
- Search.ch
PageRank and You

Get Indexed

- Submit your site to engines.
Add your URL to Google

Share your place on the net with us.

We add and update new sites to our index each time we crawl the web, and we invite you to submit your URL here. We do not add all submitted URLs to our index, and we cannot make any predictions or guarantees about when or if they will appear.

Please enter your full URL, including the http:// prefix. For example: http://www.google.com/. You may also add comments or keywords that describe the content of your page. These are used only for our information and do not affect how your page is indexed or used by Google.

Please note: Only the top-level page from a host is necessary; you do not need to submit each individual page. Our crawler, Googlebot, will be able to find the rest. Google updates its index on a regular basis, so updated or outdated link submissions are not necessary. Dead links will 'fade out' of our index on our next crawl when we update our entire index.

**URL:**

**Comments:**

**Optional: To help us distinguish between sites submitted by individuals and those automatically entered by software robots, please type the squiggly letters shown here into the box below.**

**Add URL**

Need to remove a site from Google? For more information, [click here](http://www.google.com/addurl/).
PageRank and You

Get Indexed

- Submit your site to engines.

Content

- Use metatags properly.
- Select the right keywords for your market.
Keyword Tool

The Keyword Tool generates potential keywords for your ad campaign and reports their Google statistics, including search performance and seasonal trends. Start your search by entering your own keyword phrases or a specific URL. You can then add new keywords to the green box at the right. Learn more

Important note: Please note that we cannot guarantee that these keywords will improve your campaign performance. We also reserve the right to disapprove any new keywords you add. Keep in mind that you alone are responsible for the keywords you select and for making sure that your use of the keywords does not violate any applicable laws, including any applicable trademark laws. For more details, please review our Terms and Conditions.

Results are tailored to English, United States Edit

Keyword Variations | Site-Related Keywords

Enter one keyword or phrase per line:

<table>
<thead>
<tr>
<th>Keyword Variations</th>
<th>Site-Related Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>pagerank</td>
<td>Use synonyms</td>
</tr>
</tbody>
</table>

Get More Keywords

Choose data to display: Search Volume Trends

More specific keywords - sorted by relevance

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Avg Search Volume</th>
<th>Search Volume Trends (Jan - Dec 2006)</th>
<th>Highest Volume Occurred in</th>
<th>Match Type:</th>
</tr>
</thead>
<tbody>
<tr>
<td>pagerank</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>page rank</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>search engine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>link popularity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pop up blocker</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>search engines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>backward links</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>page rank checker</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>future page</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Selected Keywords:

Click 'Sign up with these keywords' when you are finished building your keyword list.

No keywords added yet

Sign up with these keywords

https://adwords.google.com/select/KeywordToolExternal?defaultView=3
<table>
<thead>
<tr>
<th>Keyword</th>
<th>Last Update</th>
<th>Date</th>
<th>Add</th>
</tr>
</thead>
<tbody>
<tr>
<td>future page rank</td>
<td>Oct</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>check page rank</td>
<td>Oct</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>page rank tool</td>
<td>Dec</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>my page rank</td>
<td>Oct</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>page rank calculator</td>
<td>Oct</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>live page rank</td>
<td>No data</td>
<td>Jan</td>
<td>Add »</td>
</tr>
<tr>
<td>increase page rank</td>
<td>Dec</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>page rank update</td>
<td>Feb</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>page rank prediction</td>
<td>Feb</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>page rank predictor</td>
<td>No data</td>
<td>Jan</td>
<td>Add »</td>
</tr>
<tr>
<td>yahoo page rank</td>
<td>Oct</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>powered by</td>
<td>Jun</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>web page</td>
<td>Mar</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>page rank algorithm</td>
<td>No data</td>
<td>Jan</td>
<td>Add »</td>
</tr>
<tr>
<td>page rank toolbar</td>
<td>Dec</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>what is my page rank</td>
<td>No data</td>
<td>Jan</td>
<td>Add »</td>
</tr>
<tr>
<td>find page rank</td>
<td>Oct</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>high page rank</td>
<td>Jun</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>web page rank</td>
<td>Jan</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>alexa page rank</td>
<td>No data</td>
<td>Jan</td>
<td>Add »</td>
</tr>
<tr>
<td>page rank 10</td>
<td>No data</td>
<td>Jan</td>
<td>Add »</td>
</tr>
<tr>
<td>improve page rank</td>
<td>Jun</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>how to increase page rank</td>
<td>Aug</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>page rank lookup</td>
<td>Nov</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>what is page rank</td>
<td>Oct</td>
<td>Add »</td>
<td></td>
</tr>
<tr>
<td>page rank finder</td>
<td>No data</td>
<td>Jan</td>
<td>Add »</td>
</tr>
</tbody>
</table>
PageRank and You

Get Indexed
- Submit your site to engines.

Content
- Use metatags properly.
- Select the right keywords for your market.

Links
- What is your current PageRank?
- Find who links to you.
- Be wary of link exchange programs.
Advanced Search

Find results
- with all of the words
- with the exact phrase
- with at least one of the words
- without the words

Language
- Return pages written in
  - any language

File Format
- Only
  - return results of the file format
  - any format

Date
- Return web pages updated in the
  - anytime

Numeric Range
- Return web pages containing numbers between
  - and

Occurrences
- Return results where my terms occur
  - anywhere in the page

Domain
- Only
  - return results from the site or domain
  - e.g. google.com, .org
  - not filtered by license

Usage Rights
- Return results that are
  - not filtered by license

SafeSearch
- No filtering
- Filter using SafeSearch

Page-Specific Search

Similar
- Find pages similar to the page
  - e.g. www.google.com/help.html

Links
- Find pages that link to the page

Topic-Specific Searches

Google Book Search - Search the full text of books
New! Google Code Search - Search public source code
Google Scholar - Search scholarly papers
Google News archive search - Search historical news

Apple Macintosh - Search for all things Mac
BSD Unix - Search web pages about the BSD operating system
Linux - Search all penguin-friendly pages
Microsoft - Search Microsoft-related pages

U.S. Government - Search all U.S. federal, state and local government sites
Universities - Search a specific school's website
PageRank and You

Get Indexed
- Submit your site to engines.

Content
- Use metatags properly.
- Select the right keywords for your market.

Links
- What is your current PageRank?
- Find who links to you.
- Be wary of link exchange programs.

Paying Options
- Sponsored Links
- SEOs: Fortune Interactive
II. On/Off-Page Factors Relative Importance

Notice that the first metric, body word count, is the most significant metric according to the above graph. Since off-page factors have been shown again and again to be the most important metrics because they are harder to abuse than on-page factors, the above graph should not be taken to be declaring outright that body word count is the most important of all the metrics measured by SEMLogic. Rather, the body word count number is so "important" in the above graph because of the wordiness of the webpages in this particular competitive landscape. We recommend that you pay little attention to that metric, and focus in particular on the four remaining off-page metrics, Inbound Link TAI (Theme Relevancy), Inbound Link Quality, Inbound Link Title Keyword Density, and Inbound Link Anchor Keyword Density. A campaign focused primarily (although not exclusively) on those off-page metrics, with inbound link TAI in particular, should have a significantly positive impact on the quality of your web page for the keyword being optimized for Google.

The above graph also demonstrates the dominant sizes of Body Word Count (42%, but not important), IBL TAI (20%), and the remaining off-page factors.