Google’s PageRank and Beyond:
The Science of Search Engine Rankings

Amy Langville
langvillea@cofc.edu

Department of Mathematics
College of Charleston
Charleston, SC

ILAS 7/16/07
Outline

- Introduction to Information Retrieval
- Elements of a Search Engine
- Link Analysis
- Current Issues in Web Search
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   Computer
1960s         Salton’s SMART system
1989          Berner-Lee’s WWW
System for the Mechanical Analysis and Retrieval of Text

Harvard 1962 – 1965

Cornell 1965 – 1970

- Implemented on IBM 7094 & IBM 360
- Based on matrix methods

Gerard Salton
Term–Document Matrices

Start with dictionary of terms

Words or phrases (e.g., landing gear)
Term–Document Matrices

Start with dictionary of terms
Words or phrases (e.g., landing gear)

Index Each Document
Humans scour pages and mark key terms
Term–Document Matrices

Start with dictionary of terms

Words or phrases (e.g., landing gear)

Index Each Document

Humans scour pages and mark key terms

Count $f_{ij} =$ # times term $i$ appears in document $j$
**Term–Document Matrices**

Start with dictionary of terms

Words or phrases (e.g., *landing gear*)

Index Each Document

Humans scour pages and mark key terms

Count \( f_{ij} = \# \text{ times term } i \text{ appears in document } j \)

**Term–Document Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Doc 1</th>
<th>Doc 2</th>
<th>( \cdots )</th>
<th>Doc ( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Term 1</strong></td>
<td>( f_{11} )</td>
<td>( f_{12} )</td>
<td>( \cdots )</td>
<td>( f_{1n} )</td>
</tr>
<tr>
<td><strong>Term 2</strong></td>
<td>( f_{21} )</td>
<td>( f_{22} )</td>
<td>( \cdots )</td>
<td>( f_{2n} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td><strong>Term ( m )</strong></td>
<td>( f_{m1} )</td>
<td>( f_{m2} )</td>
<td>( \cdots )</td>
<td>( f_{mn} )</td>
</tr>
</tbody>
</table>

= \( A_{m \times n} \)
Query Matching

Query Vector

\[ q^T = (q_1, q_2, \ldots, q_m) \]

\[ q_i = \begin{cases} 
1 & \text{if Term } i \text{ is requested} \\
0 & \text{if not}
\end{cases} \]
Query Matching

Query Vector

\[ q^T = (q_1, q_2, \ldots, q_m) \]

\[ q_i = \begin{cases} 
1 & \text{if Term } i \text{ is requested} \\
0 & \text{if not} 
\end{cases} \]

How Close is Query to Each Document?
Query Matching

Query Vector

\[ q^T = (q_1, q_2, \ldots, q_m) \]

\[ q_i = \begin{cases} 1 & \text{if Term } i \text{ is requested} \\ 0 & \text{if not} \end{cases} \]

How Close is Query to Each Document?

i.e., how close is \( q \) to each column \( A_i \)?
Query Matching

Query Vector

\[ \mathbf{q}^T = (q_1, q_2, \ldots, q_m) \]

\[ q_i = \begin{cases} 1 & \text{if Term } i \text{ is requested} \\ 0 & \text{if not} \end{cases} \]

How Close is Query to Each Document?

i.e., how close is \( \mathbf{q} \) to each column \( \mathbf{A}_i \)?

Use \( \delta_i = \cos \theta_i = \frac{\mathbf{q}^T \mathbf{A}_i}{\|\mathbf{q}\| \|\mathbf{A}_i\|} \)
Query Matching

Query Vector

\[ \mathbf{q}^T = (q_1, q_2, \ldots, q_m) \]

\[ q_i = \begin{cases} 
1 & \text{if Term } i \text{ is requested} \\
0 & \text{if not} 
\end{cases} \]

How Close is Query to Each Document?

i.e., how close is \( \mathbf{q} \) to each column \( \mathbf{A}_i \) ?

Use \( \delta_i = \cos \theta_i = \frac{\mathbf{q}^T \mathbf{A}_i}{\| \mathbf{q} \| \| \mathbf{A}_i \|} \)

Rank documents by size of \( \delta_i \)

Return Document \( i \) to user when \( \delta_i \geq tol \)
Susan Dumais’s Improvement

- Approximate $A$ with a lower rank matrix
- Effect is to compress data in $A$

- 2 patents for Bell/Telcordia

- LATENT SEMANTIC INDEXING
Traditional IR

Pros

- Finds hidden connections
Traditional IR

Pros

• Finds hidden connections

• Can be adapted to identify document clusters
  — Text mining applications
Traditional IR

Pros

• Finds hidden connections

• Can be adapted to identify document clusters
  — Text mining applications

• Performs well on document collections that are
  ▶ Small + Homogeneous + Static
Traditional IR

Pros

- Finds hidden connections
- Can be adapted to identify document clusters
  - Text mining applications
- Performs well on document collections that are
  - Small + Homogeneous + Static

Cons

- Rankings are query dependent
  - Rank of each doc is recomputed for each query
Traditional IR

Pros

- Finds hidden connections
- Can be adapted to identify document clusters
  - Text mining applications
- Performs well on document collections that are
  - Small + Homogeneous + Static

Cons

- Rankings are query dependent
  - Rank of each doc is recomputed for each query
- Only semantic content used
  - Can be spammed + Link structure ignored
Traditional IR

Pros
• Finds hidden connections
• Can be adapted to identify document clusters
  — Text mining applications
• Performs well on document collections that are
  ▶ Small + Homogeneous + Static

Cons
• Rankings are query dependent
  — Rank of each doc is recomputed for each query
• Only semantic content used
  — Can be spammed + Link structure ignored
• Difficult to add & delete documents
Traditional IR

Pros

• Finds hidden connections
• Can be adapted to identify document clusters
  — Text mining applications
• Performs well on document collections that are
  ▶ Small + Homogeneous + Static

Cons

• Rankings are query dependent
  — Rank of each doc is recomputed for each query
• Only semantic content used
  — Can be spammed + Link structure ignored
• Difficult to add & delete documents
• Finding optimal compression requires empirical tuning
Trad. IR applied to Web

the pre-1998 Web Index

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

$k$-step transition matrix, 179

- $\alpha$ 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 convergence rate, 165
- 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
- Browne, 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
- co-citation, 123, 201
- co-reference, 123, 201
- Collatz–Wielandt formula, 168, 172
- complex networks, 30
- compressed matrix storage, 76
- condition number, 59, 71, 155
- Condorcet, 136
- connected components, 127, 133
Trad. IR applied to Web

the pre-1998 Web Index

- 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)
Trad. IR applied to Web

the pre-1998 Web Index

- 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
Sentiments about 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

- uses hyperlink structure to focus the relevant set
- combine traditional IR score with popularity score
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 - 400 X bigger than Surface Web
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 - 400 X bigger than Surface Web

• 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 - 400 X bigger than Surface Web

- 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 - 400 X bigger than Surface Web

- 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 - 400 X bigger than Surface Web

- 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

Queries

Result

Indexing Module

Indexes

Ranking Module

Indexes

Special-purpose indexes

Structure Index

Content Index
Elements of a Web Search Engine

- Crawler Module
- Page Repository
- Indexing Module
  - Indexes
    - Content Index
    - Structure Index
    - Special-purpose indexes
- Ranking Module
- User
  - Queries
  - Results
- WWW
- Query Module
- Indexes
- Query-independent
The Ranking Module (generates popularity scores)

- Measure the importance of each page
The Ranking Module (generates popularity scores)

- Measure the importance of each page

- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
Elements of a Web Search Engine

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

query-independent
The Ranking Module (generates popularity scores)

- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
- Compute these measures off-line long before any queries are processed
Elements of a Web Search Engine

WWW

Crawler Module

Page Repository

User

Indexing Module

Indexes

Ranking Module

Queries

Results

Query Module

Special-purpose indexes

Content Index

Structure Index

query-independent
The **Ranking Module** (generates popularity scores)

- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
- Compute these measures off-line long before any queries are processed
- Google’s PageRank© technology distinguishes it from all competitors
The Ranking Module (generates popularity scores)

- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
- Compute these measures off-line long before any queries are processed

- Google’s PageRank\textsuperscript{©} technology distinguishes it from all competitors

\textbf{Google’s PageRank} = Google’s $\text{$$$$}$
Take Your Pick

Amount of Internet search results that Web surfers typically scan before selecting one

A few search results*  23%
First page of search results  39%
First two pages  19%
More than first three pages  10%
First three pages  9%

*Top results without reading through the whole page
Note: Sample size is 2,369 people
Sources: JupiterResearch; iProspect
Business intelligence - Wikipedia, the free encyclopedia
Business intelligence (BI) is a business management term which refers to applications and technologies which are used to gather, provide access to, ...
en.wikipedia.org/wiki/Business_intelligence - 43k - Cached - Similar pages

Business Intelligence .com :: The Resource for Business Intelligence
The Business Intelligence resource for business and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ...
www.businessintelligence.com/ - 74k - Apr 15, 2007 - Cached - Similar pages

Business Intelligence and Performance Management Software ...
Business intelligence and business performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ...
† Stock quote for COGN
www.cognos.com/ - 32k - Cached - Similar pages

Oracle Business Intelligence Solutions
The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ...
www.oracle.com/solutions/ business_intelligence/index.html - 55k - Cached - Similar pages

Business Intelligence - Management Best Practice Reports
Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ...
www.business-intelligence.co.uk/ - 18k - Cached - Similar pages

Intelligent Enterprise: Better Insight for Business Decisions
Business intelligence - Wikipedia, the free encyclopedia
Business intelligence (BI) is a business management term which refers to applications and technologies which are used to gather, provide access to, ...
en.wikipedia.org/wiki/Business_intelligence - 43k - Cached - Similar pages

Business Intelligence .com :: The Resource for Business Intelligence
The Business Intelligence resource for business and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ...
www.businessintelligence.com/ - 74k - Apr 15, 2007 - Cached - Similar pages

Business Intelligence and Performance Management Software ... 
Business intelligence and business performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ...
Stock quote for COGN
www.cognos.com/ - 32k - Cached - Similar pages

Oracle Business Intelligence Solutions
The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ...
www.oracle.com/solutions/ business_intelligence/index.html - 55k - Cached - Similar pages

Business Intelligence - Management Best Practice Reports
Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ...
www.business-intelligence.co.uk/ - 18k - Cached - Similar pages

Business Intelligence - Management Best Practice Reports
Intelligent Enterprise: Better Insight for Business Decisions
Business intelligence (BI) is a business management term which refers to applications and technologies which are used to gather, provide access to, ... en.wikipedia.org/wiki/Business_intelligence - 43k - Cached - Similar pages

The Business Intelligence resource for business and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ... www.businessintelligence.com/ - 74k - Apr 15, 2007 - Cached - Similar pages

Business intelligence and business performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ... ☀ Stock quote for COGN www.cognos.com/ - 32k - Cached - Similar pages

The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ... www.oracle.com/solutions/business_intelligence/index.html - 55k - Cached - Similar pages

Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ... www.business-intelligence.co.uk/ - 18k - Cached - Similar pages

How To Measure “Importance”

Landmark Result Paper

Survey Paper—Big Bib
How To Measure “Importance”

Landmark Result Paper

Authorities

Survey Paper—Big Bib

Hubs
How To Measure “Importance”

- Good hubs point to good authorities
- Good authorities are pointed to by good hubs
HITS
Hypertext Induced Topic Search (1998)

Determine Authority & Hub Scores

- \( a_i = \) authority score for \( P_i \)
- \( h_i = \) hub score for \( P_i \)
HITS
Hypertext Induced Topic Search (1998)

Determine Authority & Hub Scores

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

Successive Refinement

- Start with $h_i = 1$ for all pages $P_i$  \[ h_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \]
HITS
Hypertext Induced Topic Search (1998)

Determine Authority & Hub Scores

- \( a_i \) = authority score for \( P_i \)
- \( h_i \) = hub score for \( P_i \)

Successive Refinement

- Start with \( h_i = 1 \) for all pages \( P_i \) \( \Rightarrow \) \( h_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \)
- Define Authority Scores (on the first pass)

\[
 a_i = \sum_{j: P_j \rightarrow P_i} h_j
\]
HITS

Hypertext Induced Topic Search (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 \( h_i = 1 \) for all pages \( P_i \) \( \Rightarrow \) \( h_0 = \begin{bmatrix} 1 \\ 1 \\
\vdots \\
1 \end{bmatrix} \)
- Define Authority Scores (on the first pass)

\[
a_i = \sum_{j: P_j \rightarrow P_i} h_j \quad \Rightarrow \quad a_1 = \begin{bmatrix} a_1 \\ a_2 \\
\vdots \\
a_n \end{bmatrix} = L^T h_0
\]

\[
L_{ij} = \begin{cases} 
1 & P_i \rightarrow P_j \\
0 & P_i \not\rightarrow P_j 
\end{cases}
\]
HITS Algorithm

Refine Hub Scores

- \( h_i = \sum_{j: P_i \rightarrow P_j} a_j \) \( \Rightarrow \) \( h_1 = \Lambda a_1 \)

\[
L_{ij} = \begin{cases} 
1 & P_i \rightarrow P_j \\
0 & P_i \not\rightarrow P_j 
\end{cases}
\]
HITS Algorithm

Refine Hub Scores

- \( h_i = \sum_{j: P_i \rightarrow P_j} a_j \) \( \Rightarrow \) \( h_1 = L a_1 \)

\[
L_{ij} = \begin{cases} 
1 & P_i \rightarrow P_j \\
0 & P_i \not\rightarrow P_j 
\end{cases}
\]

Successively Re-refine Authority & Hub Scores

- \( a_1 = L^T h_0 \)
HITS Algorithm

Refine Hub Scores

• \( h_i = \sum_{j : P_i \rightarrow P_j} a_j \Rightarrow h_1 = L a_1 \)

Successively Re-refine Authority & Hub Scores

• \( a_1 = L^T h_0 \)
  • \( h_1 = L a_1 \)

\[ L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \not\rightarrow P_j \end{cases} \]
HITS Algorithm

Refine Hub Scores

- \( h_i = \sum_{j: P_i \rightarrow P_j} a_j \quad \Rightarrow \quad h_1 = L a_1 \)

\[
L_{ij} = \begin{cases} 
1 & P_i \rightarrow P_j \\
0 & P_i \not\rightarrow P_j
\end{cases}
\]

Successively Re-refine Authority & Hub Scores

- \( a_1 = L^T h_0 \)
- \( h_1 = L a_1 \)
- \( a_2 = L^T h_1 \)
HITS Algorithm

Refine Hub Scores

\[ h_i = \sum_{j: P_i \to P_j} a_j \implies h_1 = L a_1 \]

Successively Re-refine Authority & Hub Scores

\[ a_1 = L^T h_0 \]
\[ h_1 = L a_1 \]
\[ a_2 = L^T h_1 \]
\[ h_2 = L a_2 \]

\[ L_{ij} = \begin{cases} 1 & P_i \to P_j \\
0 & P_i \not\to P_j \end{cases} \]
HITS Algorithm

Refine Hub Scores

\[ h_i = \sum_{j:P_i \rightarrow P_j} a_j \implies h_1 = L a_1 \]

Successively Re-refine Authority & Hub Scores

\[ a_1 = L^T h_0 \]
\[ h_1 = L a_1 \]
\[ a_2 = L^T h_1 \]
\[ h_2 = L a_2 \]

Combined Iterations

\[ A = L^T L \text{ (authority matrix)} \]
HITS Algorithm

Refine Hub Scores

- $h_i = \sum_{j:P_i \rightarrow P_j} a_j \Rightarrow h_1 = La_1$

Successively Re-refine Authority & Hub Scores

- $a_1 = L^T h_0$
  - $h_1 = La_1$
  - $a_2 = L^T h_1$
    - $h_2 = La_2$

Combined Iterations

- $A = L^T L$ (authority matrix) \quad $a_k = Aa_{k-1} \rightarrow$ e-vector (direction)
HITS Algorithm

Refine Hub Scores

- \( h_i = \sum_{j : P_i \rightarrow P_j} a_j \) \( \Rightarrow \) \( h_1 = La_1 \)

Successively Re-refine Authority & Hub Scores

- \( a_1 = L^T h_0 \)
  - \( h_1 = La_1 \)
  - \( a_2 = L^T h_1 \)
    - \( h_2 = La_2 \)
...

Combined Iterations

- \( A = L^T L \) (authority matrix) \( a_k = Aa_{k-1} \rightarrow \text{e-vector} \) (direction)
- \( H = LL^T \) (hub matrix) \( h_k = Hh_{k-1} \rightarrow \text{e-vector} \) (direction)
HITS Algorithm

Refine Hub Scores

- $h_i = \sum_{j: P_i \rightarrow P_j} a_j \quad \Rightarrow \quad h_1 = L a_1$

Successively Re-refine Authority & Hub Scores

- $a_1 = L^T h_0$
  - $h_1 = L a_1$
  - $a_2 = L^T h_1$
    - $h_2 = L a_2$
  ...

Combined Iterations

- $A = L^T L$ (authority matrix) \hspace{1cm} $a_k = A a_{k-1} \rightarrow \text{e-vector}$ (direction)
- $H = LL^T$ (hub matrix) \hspace{1cm} $h_k = H h_{k-1} \rightarrow \text{e-vector}$ (direction)

!! May not be uniquely defined if A or H is reducible !!
Compromise

1. Do direct query matching
Compromise

1. Do direct query matching

2. Build neighborhood graph
Compromise

1. Do direct query matching
2. Build neighborhood graph
3. Compute authority & hub scores for just the neighborhood
Pros & Cons

Advantages

• Returns satisfactory results
  — Client gets both authority & hub scores
Pros & Cons

Advantages

• Returns satisfactory results
  — Client gets both authority & hub scores
• Some flexibility for making refinements
Pros & Cons

Advantages

• Returns satisfactory results
  — Client gets both authority & hub scores
• Some flexibility for making refinements

Disadvantages

• Too much has to happen while client is waiting
Pros & Cons

Advantages

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

Disadvantages

- Too much has to happen while client is waiting
  - Custom built neighborhood graph needed for each query
Pros & Cons

**Advantages**

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

**Disadvantages**

- Too much has to happen while client is waiting
  - Custom built neighborhood graph needed for each query
  - Two eigenvector computations needed for each query
Pros & Cons

Advantages

• Returns satisfactory results
  — Client gets both authority & hub scores
• Some flexibility for making refinements

Disadvantages

• Too much has to happen while client is waiting
  — Custom built neighborhood graph needed for each query
  — Two eigenvector computations needed for each query
• Scores can be manipulated by creating artificial hubs
HITS Applied
The New Age of Google

The Search Giant Has Changed Our Lives. Can Anybody Catch These Guys? By Steven Levy

PLUS: The Future of Digital Voting

Google founders Larry Page and Sergey Brin
Google’s PageRank

(Lawrence Page & Sergey Brin 1998)

The Google Goals

• Create a PageRank $r(P)$ that is not query dependent
  ▶ Off-line calculations — No query time computation

• Let the Web vote with in-links
  ▶ But not by simple link counts
    — One link to $P$ from Yahoo! is important
    — Many links to $P$ from me is not

• Share The Vote
  ▶ Yahoo! casts many “votes”
    — value of vote from Yahoo! is diluted
  ▶ If Yahoo! “votes” for $n$ pages
    — Then $P$ receives only $r(Y)/n$ credit from $Y$
Google’s PageRank

(Lawrence Page & Sergey Brin 1998)

The Google Goals

- Create a PageRank $r(P)$ that is not query dependent
  - Off-line calculations — No query time computation

- Let the Web vote with in-links
  - But not by simple link counts
    - One link to $P$ from Yahoo! is important
    - Many links to $P$ from me is not

- Share The Vote
  - Yahoo! casts many “votes”
    - value of vote from Yahoo! is diluted
  - If Yahoo! “votes” for $n$ pages
    - Then $P$ receives only $r(Y)/n$ credit from $Y$
Google's PageRank

(Lawrence Page & Sergey Brin 1998)

The Google Goals

- Create a PageRank \( r(P) \) that is not query dependent
  - Off-line calculations — No query time computation
- Let the Web vote with in-links
  - But not by simple link counts
    - One link to \( P \) from Yahoo! is important
    - Many links to \( P \) from me is not
- Share The Vote
  - Yahoo! casts many “votes”
    - value of vote from \( Yahoo! \) is diluted
  - If Yahoo! “votes” for \( n \) pages
    - Then \( P \) receives only \( r(Y)/n \) credit from \( Y \)
Google’s PageRank

(Lawrence Page & Sergey Brin 1998)

The Google Goals

- Create a PageRank $r(P)$ that is not query dependent
  - Off-line calculations — No query time computation
- Let the Web vote with in-links
  - But not by simple link counts
    - One link to $P$ from Yahoo! is important
    - Many links to $P$ from me is not
- Share The Vote
  - Yahoo! casts many “votes”
    - value of vote from Yahoo! is diluted
  - If Yahoo! “votes” for $n$ pages
    - Then $P$ receives only $r(Y)/n$ credit from $Y$
PageRank

The Definition

\[ r(P) = \sum_{P \in B_P} \frac{r(P)}{|P|} \]

\[ B_P = \{ \text{all pages pointing to } P \} \]
\[ |P| = \text{number of out links from } P \]
PageRank

The Definition

\[ r(P) = \sum_{P \in \mathcal{B}_P} \frac{r(P)}{|P|} \]

where \( \mathcal{B}_P = \{ \text{all pages pointing to } P \} \) and \( |P| = \text{number of out links from } P \).

Successive Refinement

Start with \( r_0(P_i) = \frac{1}{n} \) for all pages \( P_1, P_2, \ldots, P_n \).
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

\[ r_1(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_0(P)}{|P|} \]
PageRank

The Definition

\[ r(P) = \sum_{P \in B_P} \frac{r(P)}{|P|} \]

\( B_P = \{ \text{all pages pointing to } P \} \)

\( |P| = \text{number of out links from } P \)

Successive Refinement

Start with \( r_0(P_i) = 1/n \) for all pages \( P_1, P_2, \ldots, P_n \)

Iteratively refine rankings for each page

\[ r_1(P_i) = \sum_{P \in B_{P_i}} \frac{r_0(P)}{|P|} \]

\[ r_2(P_i) = \sum_{P \in B_{P_i}} \frac{r_1(P)}{|P|} \]
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 \( r_0(P_i) = \frac{1}{n} \) for all pages \( P_1, P_2, \ldots, P_n \)

Iteratively refine rankings for each page

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

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

\[ \vdots \]

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

After Step $k$

$\pi_k^T = [r_k(P_1), r_k(P_2), \cdots, r_k(P_n)]$
In Matrix Notation

After Step $k$

\[ \pi_k^T = [r_k(P_1), r_k(P_2), \ldots, r_k(P_n)] \]

\[ \pi_{k+1}^T = \pi_k^T H \quad \text{where} \quad h_{ij} = \begin{cases} \frac{1}{|P_i|} & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases} \]
In Matrix Notation

After Step $k$

$\pi^T_k = [r_k(P_1), r_k(P_2), \cdots, r_k(P_n)]$

$\pi^T_{k+1} = \pi^T_k H$ where $h_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$

PageRank vector $= \pi^T = \lim_{k \rightarrow \infty} \pi^T_k$ is an eigenvector for $H$

Provided that the limit exists
$$H = \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ P_2 & P_3 & P_4 & P_5 & P_6 & \end{pmatrix}$$
Tiny Web

\[
H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0
\end{pmatrix}
\]
$H = \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ P_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ P_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ P_4 & 0 & 0 & 0 & 0 & 0 & 0 \\ P_5 & 0 & 0 & 0 & 0 & 0 & 0 \\ P_6 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
Tiny Web

\[ H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
P_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
\end{pmatrix} \]
$H = \begin{pmatrix}
  P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
  P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
  P_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
  P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
  P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
  P_5 & 0 & 0 & 0 & 0 & 0 & 0 \\
  P_6 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$
Tiny Web

\[
H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2
\end{pmatrix}
\]
Tiny Web

\[ H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0 
\end{pmatrix} \]
Tiny Web

A random walk on the Web Graph

\[
H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
P_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
P_6 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}
\]
A random walk on the Web Graph

PageRank = $\pi_i$ = amount of time spent at $P_i$
Tiny Web

\[ H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix} \]

▷ A random walk on the Web Graph

▷ PageRank = \( \pi_i \) = amount of time spent at \( P_i \)

▷ Dead end page (nothing to click on) — a “dangling node”
### Tiny Web

A random walk on the Web Graph

PageRank = $\pi_i$ = amount of time spent at $P_i$

Dead end page (nothing to click on) — a “dangling node”

$\pi^T = (0, 1, 0, 0, 0, 0) = e$-vector  $\Rightarrow$ Page $P_2$ is a “rank sink”

![Web Graph](image)

$$H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}$$
The Fix

Allow Web Surfers To Make Random Jumps
The Fix

Allow Web Surfers To Make Random Jumps

Replace zero rows with \( \frac{e^T}{n} = \left( \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right) \)

\[
S = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
P_2 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
P_6 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}
\]
The Fix

Allow Web Surfers To Make Random Jumps

— Replace zero rows with $\frac{e^T}{n} = \left( \frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n} \right)$

$$S = \begin{pmatrix}
    P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
    P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
    P_2 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
    P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
    P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
    P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
    P_6 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}$$

— $S = H + \frac{ae^T}{6}$ is now row stochastic $\Rightarrow \rho(S) = 1$
The Fix

Allow Web Surfers To Make Random Jumps

— Replace zero rows with

\[ \mathbf{e}^T = \left( \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right) \]

\[
\mathbf{S} = \\
\begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
P_2 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
P_6 & 0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}
\]

— \[ \mathbf{S} = \mathbf{H} + \frac{\mathbf{a} \mathbf{e}^T}{6} \] is now row stochastic \[ \rho(\mathbf{S}) = 1 \]

— Perron says \[ \exists \quad \mathbf{\pi}^T \geq 0 \] s.t. \[ \mathbf{\pi}^T = \mathbf{\pi}^T \mathbf{S} \] with \[ \sum_i \pi_i = 1 \]
Nasty Problem

The Web Is Not Strongly Connected
Nasty Problem

The Web Is Not Strongly Connected

- S is reducible

\[ S = \begin{pmatrix}
    P_1 & P_2 & P_3 & & P_4 & P_5 & P_6 \\
    P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
    P_2 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
    P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
    P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
    P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
    P_6 & 0 & 0 & 0 & 1 & 0 & 0 
\end{pmatrix} \]
Nasty Problem

The Web Is Not Strongly Connected

- S is reducible

\[
S = \begin{pmatrix}
P_1 & P_2 & P_3 & & & \\
P_1 & 0 & 1/2 & 1/2 & & \\
P_2 & 1/6 & 1/6 & 1/6 & & \\
P_3 & 1/3 & 1/3 & 0 & & \\
P_4 & 0 & 0 & 0 & & \\
P_5 & 0 & 0 & 0 & & \\
P_6 & 0 & 0 & 0 & & \\
\end{pmatrix}
\]

- Reducible \(\implies\) PageRank vector is not well defined

- Frobenius says S needs to be \textit{irreducible} to ensure a unique \(\pi^T > 0\) s.t. \(\pi^T = \pi^T S\) with \(\sum_i \pi_i = 1\)
Irreducibility Is Not Enough

Could Get Trapped Into A Cycle \((P_i \rightarrow P_j \rightarrow P_i)\)
Irreducibility Is Not Enough

Could Get Trapped Into A Cycle \((P_i \rightarrow P_j \rightarrow P_i)\)

- The powers \(S^k\) fail to converge
Irreducibility Is Not Enough

Could Get Trapped Into A Cycle \((P_i \rightarrow P_j \rightarrow P_i)\)

— The powers \(S^k\) fail to converge

— \(\pi_{k+1}^T = \pi_k^T S\) fails to convergence
Irreducibility Is Not Enough

Could Get Trapped Into A Cycle \((P_i \rightarrow P_j \rightarrow P_i)\)

- The powers \(S^k\) fail to converge
- \(\pi_{k+1}^T = \pi_k^T S\) fails to converge

Convergence Requirement

- Perron–Frobenius requires \(S\) to be primitive
Irreducibility Is Not Enough

Could Get Trapped Into A Cycle \((P_i \rightarrow P_j \rightarrow P_i)\)

— The powers \(S^k\) fail to converge

— \(\pi_{k+1}^T = \pi_k^T S\) fails to convergence

Convergence Requirement

— Perron–Frobenius requires \(S\) to be primitive

— No eigenvalues other than \(\lambda = 1\) on unit circle
Irreducibility Is Not Enough

Could Get Trapped Into A Cycle \((P_i \rightarrow P_j \rightarrow P_i)\)

— The powers \(S^k\) fail to converge

— \(\pi_{k+1}^T = \pi_k^T S\) fails to convergence

Convergence Requirement

— Perron–Frobenius requires \(S\) to be primitive

— No eigenvalues other than \(\lambda = 1\) on unit circle

— Frobenius proved \(S\) is primitive \(\iff S^k > 0\) for some \(k\)
The Google Fix

Allow A Random Jump From Any Page

\[ G = \alpha S + (1 - \alpha)E > 0, \quad E = ee^T/n, \quad 0 < \alpha < 1 \]
The Google Fix

Allow A Random Jump From Any Page

\[ G = \alpha S + (1 - \alpha)E > 0, \quad E = ee^T/n, \quad 0 < \alpha < 1 \]

\[ G = \alpha H + uv^T > 0 \quad u = \alpha a + (1 - \alpha)e, \quad v^T = e^T/n \]
The Google Fix

Allow A Random Jump From Any Page

\[ G = \alpha S + (1 - \alpha)E > 0, \quad E = ee^T/n, \quad 0 < \alpha < 1 \]

\[ G = \alpha H + uv^T > 0 \quad \quad u = \alpha a + (1 - \alpha)e, \quad v^T = e^T/n \]

PageRank vector \[ \pi^T = \text{left-hand Perron vector of } G \]
The Google Fix

Allow A Random Jump From Any Page

\[ G = \alpha S + (1 - \alpha)E > 0, \quad E = ee^T/n, \quad 0 < \alpha < 1 \]

\[ G = \alpha H + uv^T > 0 \quad \quad \quad u = \alpha a + (1 - \alpha)e, \quad v^T = e^T/n \]

\[ \text{PageRank vector} \quad \pi^T = \text{left-hand Perron vector of } G \]

Some Happy Accidents

\[ x^T G = \alpha x^T H + \beta v^T \quad \text{Sparse computations with the original link structure} \]
The Google Fix

Allow A Random Jump From Any Page

$G = \alpha S + (1 - \alpha)E > 0, \quad E = ee^T/n, \quad 0 < \alpha < 1$

$G = \alpha H + uv^T > 0 \quad u = \alpha a + (1 - \alpha)e, \quad v^T = e^T/n$

PageRank vector $\pi^T = \text{left-hand Perron vector of } G$

Some Happy Accidents

$x^T G = \alpha x^T H + \beta v^T \quad \text{Sparse computations with the original link structure}$

$\lambda_2(G) = \alpha \quad \text{Convergence rate controllable by Google engineers}$
The Google Fix

Allow A Random Jump From Any Page

\[ G = \alpha S + (1 - \alpha)E > 0, \quad E = ee^T/n, \quad 0 < \alpha < 1 \]

\[ G = \alpha H + uv^T > 0 \quad u = \alpha a + (1 - \alpha)e, \quad v^T = e^T/n \]

PageRank vector \[ \pi^T = \text{left-hand Perron vector of } G \]

Some Happy Accidents

\[ x^T G = \alpha x^T H + \beta v^T \quad \text{Sparse computations with the original link structure} \]

\[ \lambda_2(G) = \alpha \quad \text{Convergence rate controllable by Google engineers} \]

\[ v^T \text{ can be any positive probability vector in } G = \alpha H + uv^T \]
The Google Fix

Allow A Random Jump From Any Page

\[ G = \alpha S + (1 - \alpha)E > 0, \quad E = ee^T/n, \quad 0 < \alpha < 1 \]

\[ G = \alpha H + uv^T > 0 \quad u = \alpha a + (1 - \alpha)e, \quad v^T = e^T/n \]

PageRank vector \[ \pi^T = \text{left-hand Perron vector of } G \]

Some Happy Accidents

\[ x^T G = \alpha x^T H + \beta v^T \quad \text{Sparse computations with the original link structure} \]

\[ \lambda_2(G) = \alpha \quad \text{Convergence rate controllable by Google engineers} \]

\[ v^T \] can be any positive probability vector in \[ G = \alpha H + uv^T \]

The choice of \[ v^T \] allows for personalization
Search Issues

Spamming

- Link Farms
What’s News—

Business and Finance

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

(Article on Page A3)

The SEC signaled it may file civil charges against Morgan Stanley, alleging it doled out IPO shares based partly on investors’ commitments to buy more stock.

(Article on Page C1)

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 horse-trading 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 I and Pages A4 and A6)

Turkey’s parliament debated legislation to let the U.S. deploy 52,000 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.

(Article 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 expectations the Bush administration will lower corporate taxes.

Web Master

As the Web spreads...

Total Internet users, by household, in millions

Google’s U.S. presence expands

Top search engines, in millions of unique visitors

Yahoo Search

39.4

MSN Search

38.8

AOL Search

22.0

Ask Jeeves

13.3

Overture

6.4

Source: Forrester Research; Nielsen NetRatings

Top shopping-referral sites, in millions of referrals

Google

12.61

DeaTime

2.50

RinRate

1.93

Overture

1.04

Epinions

0.78

CNET

0.76

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.

Cat and Mouse

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

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

Exotic leatherwear Gets Cut Off

By Michael Totty

And Myylene 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 computer. 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
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 is offering its clients a chance to boost their own Google rankings by buying ads on more popular sites. SearchKing filed suit against the company in federal court in Oklahoma, claiming that Google "purposefully devalued" SearchKing and its customers, damaging its reputation and hurting its business.

Google won't comment on the case. 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 graduates Sergey Brin and Larry Page, says Web companies that want to rank highly should concentrate on improving their Web pages rather than gaming the system. "When people try to take shortcuts into their own hands, that turns into a worse experience for users," says Matt Cutts, a Google software engineer.

Coden 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 widely used 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 coding deceived search engines, which interpreted 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, said that in an internal 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 Google

The big search engines determine the laws of how commerce runs," says Mr. Massa.

"...the key to success is not just about ranking high on search engines, but also about being able to convert those visits into sales..."
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)"


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

RELATED INTERNET LINKS:
White House
Google bombing

The BBC is not responsible for the content of external internet sites

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
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 to frequent search times, 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 those 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.
Google Search: miserable failure

miserable failure
Google Search

Advanced Search Preferences Language Tools Search Tips

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. Newday 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 cucked 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
Search Issues

Spamming

- Link Farms
- Google Bombs

Personalization

- Google’s psearch, A9, Kartoo
Search Engines Seek to Get Inside Your Head

Google, Others Start to Comb Users’ Online Habits to Tailor Results to Personal Interests

By JESSICA E. VASCCELLARO
And KEVIN J. DELANEY

Search engines have long generated the same results for queries whether the person searching was a mom, mathematician or movie star. Now, who you are and what you’re interested in is starting to affect the outcome of your search.

Google Inc. and a wide range of start-ups are trying to translate factors like where you live, the ads you click on and the types of restaurants you search for into more-relevant search results. A chef who searched for “beef,” for example, might be more likely to find recipes than encyclopedia entries about livestock. And a film buff who searched for a new movie might see detailed articles about the making of the film, rather than ticket-buying sites.

Google has been enhancing and more widely deploying its search-personalization technology. Within coming weeks, Google users who are logged in will begin having their search results reordered based on information they have provided to Google. For instance, they may have entered a city to receive weather forecasts on a personalized Google home page. As a result, a user in New York who types in “Giants” might see higher search results for the football team than a user in San Francisco, who might be more interested in the Giants baseball team.

Consumers who use its Web-history service to track previous search queries currently get results that are influenced by those queries and the sites they have clicked on. The company plans eventually to offer personalization based on a user’s Web-browsing history—including sites people visited without going through Google—when users agree to let Google track it.

Also, within three to five years, Google will...
Search Issues

Spamming
- Link Farms
- Google Bombs

Personalization
- Google’s psearch, A9, Kartoo

Privacy
- 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 “nimb 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
Department over a subpoena for some of its search data. The company successfully fended off the agency’s demand in court, but several other search companies, including AOL, complied. The Justice Department sought the information to help it defend a challenge to a law that is meant to shield children from sexually explicit material.

“We supported Google at the time,” Mr. Rotenberg said, “but we also said that it was a mistake for Google to be saving so much information because it creates a risk.”

Ms. Arnold, who agreed to discuss her searches with a reporter, said she was shocked to hear that AOL had saved and published three months’ worth of them. “My goodness, it’s my whole personal life,” she said. “I had no idea somebody was looking over my shoulder.”

In the privacy of her four-bedroom home, Ms. Arnold searched for the answers to scores of life’s questions, big and small. How could she buy “school supplies for Iraq children”? What is the “safest place to live”? What is “the best season to visit Italy”?

Her searches are a catalog of intentions, curiosity, anxieties and quotidian questions. There was the day in May, for example, when she typed in “termites,” then “tea for good health” then “mature living,” all within a few hours.

Her queries mirror millions of those captured in AOL’s database, which reveal the concerns of expectant mothers, cancer patients, college students and music lovers. User No. 2178 searches for “foods to avoid when breast feeding.” No. 3482401 seeks guidance on “calorie counting.” No. 3483689 searches for the songs “Time After Time” and “Wind Beneath My Wings.”

At times, the searches appear to betray intimate emotions and personal dilemmas. No. 3505202 asks about “depression and medical leave.” No. 7268042 types “fear that spouse contemplating cheating.”

There are also many thousands of sexual queries, along with searches about “child porn” and “how to kill oneself by natural gas” that raise questions about what legal authorities can and should do with such information.

But while these searches can tell the casual observer — or the sociologist or the marketer — much about the person who typed them, they can also prove highly misleading.

At first glance, it might appear that Ms. Arnold fears she is suffering from a wide range of ailments. Her search history includes “hand tremors,” “nicotine effects on the body,” “dry mouth” and “bipolar.” But in an interview, Ms. Arnold said she routinely researched medical conditions for her friends to assuage their anxieties. Explaining her queries about nicotine, for example, she said: “I have a friend who needs to quit smoking and I want to help her do it.”

Saul Hansell contributed reporting for this article.
Search Issues

Spamming
- Link Farms
- Google Bombs

Personalization
- Google’s psearch, A9, Kartoo

Privacy
- AOL Data Leak

Data Fusion
- Search.ch
Conclusion

- Link-based scores (PageRank, HITS, etc.) are combined with content scores for final rankings.
- Link analysis has dramatically improved search.
- Many continuing CSC and MATH challenges.
Web Graphs

CSC and MATH challenges (problems of scale!)

- store adjacency matrix
- update adjacency matrix
- visualize web graph
- locate clusters in graph
Conclusion

- Link-based scores (PageRank, HITS, etc.) are combined with content scores for final rankings.
- Link analysis has dramatically improved search.
- Many continuing CSC and MATH challenges.
- The constant battle between search engines and SEOs means that companies and algorithms must adapt and innovate.
Conclusion

- Link-based scores (PageRank, HITS, etc.) are combined with content scores for final rankings.
- Link analysis has dramatically improved search.
- Many continuing CSC and MATH challenges.
- The constant battle between search engines and SEOs means that companies and algorithms must adapt and innovate.

Elegant and Exciting Application of Linear Algebra
That is Changing the World