Google’s PageRank
and Beyond

Carl Meyer

Department of Mathematics
North Carolina State University
Raleigh, NC
Beautiful mathematics eventually tends to be useful, and useful mathematics eventually tends to be beautiful.
Short History of IR

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

B. C.                                   cave paintings
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
                                            Computer
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 $f_{ij} = \# \text{ times term } i \text{ appears in document } j$
**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**

$$
\begin{pmatrix}
  f_{11} & f_{12} & \cdots & f_{1n} \\
  f_{21} & f_{22} & \cdots & f_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  f_{m1} & f_{m2} & \cdots & f_{mn}
\end{pmatrix} = A_{m \times n}
$$
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

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

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

How Close is Query to Each Document?

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

Use

\[ \delta_i = \cos \theta_i = \frac{q^T A_i}{\|q\| \|A_i\|} \]
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 \( A_i \)?

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
Latent Semantic Indexing

Use a Fourier expansion of $A$

$$A = \sum_{i=1}^{r} \sigma_{i} Z_{i}, \quad \langle Z_{i} | Z_{j} \rangle = \begin{cases} 1 & i=j, \\ 0 & i \neq j, \end{cases} \quad |\sigma_{1}| \geq |\sigma_{2}| \geq \cdots \geq |\sigma_{r}|$$

$$|\sigma_{i}| = |\langle Z_{i} | A \rangle| = \text{amount of } A \text{ in direction of } Z_{i}$$
Latent Semantic Indexing

Use a Fourier expansion of $A$

$$A = \sum_{i=1}^{r} \sigma_i Z_i,$$

$$\langle Z_i | Z_j \rangle = \begin{cases} 1 & i = j, \\ 0 & i \neq j, \end{cases}$$

$$|\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_r|$$

$$|\sigma_i| = |\langle Z_i | A \rangle| = \text{amount of } A \text{ in direction of } Z_i$$

Realign data along dominant directions $\{Z_1, \ldots, Z_k, Z_{k+1}, \ldots, Z_r\}$

— Project $A$ onto $\text{span} \{Z_1, Z_2, \cdots, Z_k\}$
Latent Semantic Indexing

Use a Fourier expansion of $A$

$$A = \sum_{i=1}^{r} \sigma_i Z_i,$$

$$\langle Z_i | Z_j \rangle = \begin{cases} 1 & i=j, \\ 0 & i \neq j, \end{cases} \quad |\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_r|$$

$$|\sigma_i| = |\langle Z_i | A \rangle| = \text{amount of } A \text{ in direction of } Z_i$$

Realign data along dominant directions $\{Z_1, \ldots, Z_k, Z_{k+1}, \ldots, Z_r\}$

— Project $A$ onto $\text{span} \{Z_1, Z_2, \cdots, Z_k\}$

**Truncate:** $A_k = P(A) = \sigma_1 Z_1 + \sigma_2 Z_2 + \cdots + \sigma_k Z_k$
Latent Semantic Indexing

Use a Fourier expansion of $A$

$$A = \sum_{i=1}^{r} \sigma_i z_i,$$

$$\langle z_i | z_j \rangle = \begin{cases} 1 & i=j, \\ 0 & i \neq j, \end{cases} \quad |\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_r|$$

$$|\sigma_i| = |\langle z_i | A \rangle| = \text{amount of } A \text{ in direction of } z_i$$

Realign data along dominant directions $\{z_1, \ldots, z_k, z_{k+1}, \ldots, z_r\}$

— Project $A$ onto $\text{span} \{z_1, z_2, \cdots, z_k\}$

Truncate: $A_k = P(A) = \sigma_1 z_1 + \sigma_2 z_2 + \cdots + \sigma_k z_k$

LSI: Query matching with $A_k$ in place of $A$

— $Doc_2$ forced closer to $Doc_1 \implies$ better chance of finding $Doc_2$
Latent Semantic Indexing

Use a Fourier expansion of $A$

$$A = \sum_{i=1}^{r} \sigma_i Z_i,$$

$$\langle Z_i | Z_j \rangle = \begin{cases} 
1 & i=j, \\
0 & i \neq j, 
\end{cases} \quad |\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_r|$$

$$|\sigma_i| = |\langle Z_i | A \rangle| = \text{amount of } A \text{ in direction of } Z_i$$

Realign data along dominant directions $\{Z_1, \ldots, Z_k, Z_{k+1}, \ldots, Z_r\}$

— Project $A$ onto $\text{span} \{Z_1, Z_2, \cdots, Z_k\}$

Truncate: $A_k = P(A) = \sigma_1 Z_1 + \sigma_2 Z_2 + \cdots + \sigma_k Z_k$

LSI: Query matching with $A_k$ in place of $A$

— $Doc_2$ forced closer to $Doc_1 \implies$ better chance of finding $Doc_2$

"Best" mathematical solution

— SVD: $A = U D V^T = \sum \sigma_i u_i v_i^T$

$$Z_i = u_i v_i^T$$
Strengths & Weaknesses

Pros

- Finds hidden connections
Strengths & Weaknesses

Pros

• Finds hidden connections

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

Pros

• Finds hidden connections

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

• Performs well on document collections that are
  ▶ Small + Homogeneous + Static
Strengths & Weaknesses

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
Strengths & Weaknesses

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
Strengths & Weaknesses

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
Strengths & Weaknesses

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
Web Stats

Different from other document collections

• It’s huge
  – Over 10 billion pages, where average page size $\approx 500\text{KB}$
  – 20 times size of Library of Congress print collection
  – Deep Web $\approx 550$ billion pages
Web Stats

Different from other document collections

- **It’s huge**
  - Over 10 billion pages, where average page size $\approx 500$KB
  - 20 times size of Library of Congress print collection
  - Deep Web $\approx 550$ billion pages

- **It’s dynamic**
  - 40% of all pages change in a week
  - 23% of .com pages change daily
  - Billions of pages added each year
Web Stats

Different from other document collections

- **It’s huge**
  - Over 10 billion pages, where average page size $\approx 500$KB
  - 20 times size of Library of Congress print collection
  - Deep Web $\approx 550$ billion pages

- **It’s dynamic**
  - 40% of all pages change in a week
  - 23% of .com pages change daily
  - Billions of pages added each year

- **It’s self-organized**
  - No standards, review process, formats
  - Errors, falsehoods, link rot, and spammers!
Web Stats

Different from other document collections

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

- **It’s dynamic**
  - 40% of all pages change in a week
  - 23% of .com pages change daily
  - Billions of pages added each year

- **It’s self-organized**
  - No standards, review process, formats
  - Errors, falsehoods, link rot, and spammers!

- **It has many users**
  - Google alone processes more than 200 million queries per day
  - Approximately 0.25 sec per query involving thousands of computers
Web Search Components

Web Crawlers

Software robots gather web pages
Web Search Components

- **Web Crawlers**: Software robots gather web pages.
- **Doc Server**: Stores docs and snippits.
Web Search Components

Web Crawlers
Software robots gather web pages

Doc Server
Stores docs and snippits

Index Server
Scans pages and does term indexing
Terms → Pages (similar to book index)
The Ranking Module

- Measure the importance of each page
The Ranking Module

- 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
The Ranking Module

- 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
The Ranking Module

- 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

- 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

Google’s PageRank = Google’s $$$$$$
The Process
The Process

query

Web Server

Index Server
The Process
Take Your Pick

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

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

*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

SQL Database Management
Enterprise Data Mgmt Solutions From Dell™. Find Out More Here www.dell.com

Business Intelligence
See what business intelligence can do for you (free interactive demo). www.InformationBuilders.com

MCITP: BI Cert Boot Camp
9-Day MCITP Certification Boot Camp Business Intelligence All Inclusive www.mcseclasses.com

Business Intelligence
Improve information integrity with real-time data integration software www.DataMirror.com

Love Data?
Empower yourself with MS BI Tools via SetFocus' Master's Program www.SetFocus.com

Business Intelligence
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
Yahoo Ad System Fails to Lift Net

Revenue Growth Declines; Project Benefits Are Seen Ramping Up in 2nd Period

By Kevin J. Delaney

Yahoo Inc. recently overhauled its online advertising system, giving some investors hope for a positive earnings surprise. So far, that hope hasn’t materialized.

The Sunnyvale, Calif., company reported an 11% drop in first-quarter profit as its revenue growth rate continued a steady decline. Yahoo’s shares fell about 8% in after-hours trading.

Some investors had raised hopes for the company’s first-quarter results following a major overhaul of Yahoo’s online advertising system dubbed Project Panama that was rolled out in recent months. But Yahoo’s revenue was in line with its earlier projection, and it stuck to its outlook for the year. The company reiterated earlier predictions that financial benefits from Panama, which includes big changes to its search-ad system designed to boost Yahoo revenue, will start kicking in during the second quarter.

Analysts said the first quarter had been expected to be a tough one when compared with earlier quarters, with benefits from Panama not yet arriving and increased competition for the graphical display advertising that some estimate represents about one-third of Yahoo’s revenue. In addition, the first quarter of last year included revenue from ad brokerage for Microsoft Corp., which has since been discontinued, making for tougher comparisons.

When commissions paid to marketing partners were factored out, Yahoo reported revenue of $1.18 billion for the first quarter, in line with its projection of $1.12 billion to $1.23 billion. Yahoo stuck to its prediction of 2007 revenue on that basis of $4.95 billion to $5.45 billion.

Yahoo reported its results after regular trading hours. In 4 p.m. Nasdaq Stock Market composite trading, shares were up 48 cents to $32.09. That is about 25% higher than their level at the beginning of the year and 4% above 12 months earlier. In after-hours trading, Yahoo shares fell about 8% to $29.51.

“People were expecting a possibility of upward guidance and we didn’t get that so the stock is giving back some of its recent gains,” said Rob Sanderson, an analyst at American Technology Research. “This should be the toughest quarter; that was the expectation going in.”

Revenue growth continued to decline at Yahoo. Revenue rose 7% in the first quarter, compared with 13% in the fourth quarter, 19% growth in the third quarter and 26% in the second.

Rival Yahoo’s earnings fall

THE ASSOCIATED PRESS

SAN FRANCISCO — Google’s first-quarter profit rose 69 percent, maintaining the online search leader’s penchant for obliterating analyst estimates.

The stellar results released Thursday left little doubt that Google has widened its lead over its closest rival in Internet search and advertising, Yahoo, whose first-quarter earnings eroded.

Google detailed its sparkling performance on the same day that several major U.S. newspaper companies announced another quarter of financial decay, underscoring an advertising shift that is enriching Internet upstarts at the expense of traditional media outlets.

Born less than a decade ago, Google now reigns as the most profitable — and probably most powerful — force on the Web.

In the latest demonstration of its clout, Google earned $1 billion, or $3.15 cents per share, during the first three months of the year. That compared with net income of $592.3 million, or $1.05 per share, in the same period last year. It was also the second consecutive quarter in which Google earned $1 billion — nearly as much money as the nation’s largest newspaper publisher, Gannett, made all of last year.

If not for expenses incurred for employee stock compensation, Google would have earned $3.68 per share.

Quarterly revenue reached a new company high of $3.66 billion, a 53 percent increase. After subtracting advertising commissions and other payments to its partners, Google’s revenue totaled $2.53 billion.

Pleasant earnings surprises have become routine for Google, which has succeeded in beating analyst estimates in all but one of 11 quarters since its ballyhooed initial public offering of stock in August 2004.

That track record had helped elevate Google’s market value to nearly $150 billion, even before the stock price surged $12.55 in Thursday’s extended trading.

As usual, Google’s financial power flowed from its search engine. That ubiquitous tool has become the hub of the Web’s largest marketing network and appears to be getting even better at identifying the right ads to show with its search results, helping elicit more revenue-generating clicks. The paid clicks on the ads within Google’s vast network increased 52 percent for the first quarter, compared with year-ago levels. And more of the clicks are occurring on Google’s own Web sites, increasing the company’s profits because the revenue doesn’t have to be shared with an advertising partner.

Although Google has been trying to develop revenue beyond the Internet, online advertising continues to produce virtually all of its profit. The company is expected to become even more dominant in that business with last year’s $1.7 billion acquisition of online video leader YouTube and its recently announced $3.1 billion deal to buy Internet ad distributor DoubleClick.

Although Google still isn’t making money from YouTube, the site is “going gangbusters,” co-founder Larry Page said during Thursday’s conference call.

Besides buying other companies, Google is investing heavily to accommodate its growth by hiring workers and adding computer capacity at its data centers. The company spent $597 million on capital expenditures in the first quarter and hired 1,506 employees to expand its workforce by 12,238 people.

Even so, the company ended the quarter with $11.9 billion in cash.
How To Measure “Importance”

Landmark Result Paper

Survey Paper—Big Bib
How To Measure “Importance”

Landmark Result Paper

Survey Paper—Big Bib

Authorities

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

- $a_i = \text{authority score for } P_i$
- $h_i = \text{hub score for } P_i$
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 $P_i$  \(\Rightarrow\)  $h_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
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)

$$a_i = \sum_{j : P_j \rightarrow P_i} h_j$$
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)

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

$$L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \nrightarrow P_j \end{cases}$$
HITS Algorithm

Refine Hub Scores

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

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

Refine Hub Scores

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

\[ L_{ij} = \begin{cases} 
1 & \text{if } P_i \rightarrow P_j \\
0 & \text{otherwise}
\end{cases} \]

Successively Re-refine Authority & Hub Scores

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

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

- \( a_1 = L^T h_0 \)
  - \( h_1 = La_1 \)

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

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

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

Refine Hub Scores

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

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

Successively Re-refine Authority & Hub Scores

\[ a_1 = L^T h_0 \]

\[ h_1 = L a_1 \]

\[ a_2 = L^T h_1 \]

\[ h_2 = L a_2 \]

...
HITS Algorithm

Refine Hub Scores

- \[ h_i = \sum_{j: P_i \to P_j} a_j \Rightarrow h_1 = La_1 \]

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

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

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

Refine Hub Scores
\[ h_i = \sum_{j: P_i \to P_j} a_j \implies h_1 = La_1 \]
\[ L_{ij} = \begin{cases} 1 & P_i \to P_j \\ 0 & P_i \nrightarrow P_j \end{cases} \]

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)
\[ a_k = A a_{k-1} \rightarrow \text{e-vector (direction)} \]
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)

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

Refine Hub Scores

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

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

Successively Re-refine Authority & Hub Scores

• \( a_1 = L^T h_0 \)
  • \( h_1 = La_1 \)

• \( a_2 = L^T h_1 \)
  • \( h_2 = La_2 \)

Combined Iterations

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

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

\( \mathcal{B}_P = \{ \text{all pages pointing to } P \} \)

\(|P| = \text{number of out links from } P\)
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 \( P_1, P_2, \ldots, P_n \)
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

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

Iteratively refine rankings for each page

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

\[ r_2(P_i) = \sum_{P \in B_{P_i}} \frac{r_1(P)}{|P|} \]
PageRank

The Definition

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

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

Successive Refinement

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

Iteratively refine rankings for each page

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

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

\[ \vdots \]

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

After Step $k$

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

After Step $k$

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

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

After Step $k$

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

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

PageRank vector $= \pi^T = \lim_{k \to \infty} \pi^T_k = \text{eigenvector for } H$

Provided that the limit exists
Tiny Web

\[
H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
P_3 & P_4 & P_5 & P_6 & & \\
P_4 & P_5 & P_6 & & & \\
P_5 & P_6 & & & & \\
P_6 & & & & &
\end{pmatrix}
\]
Tiny Web

$$H = \begin{pmatrix}
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$
Tiny Web

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

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

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

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

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

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

▷ A random walk on the Web Graph
Tiny Web

A random walk on the Web Graph

PageRank = $\pi_i = \text{amount of time spent at } P_i$
A random walk on the Web Graph

PageRank = $\pi_i = \text{amount of time spent at } P_i$

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

Allow Web Surfers To Make Random Jumps
The Fix

Allow Web Surfers To Make Random Jumps

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

\[
\mathbf{s} = \begin{pmatrix}
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}
\]
The Fix

Allow Web Surfers To Make Random Jumps

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

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

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

Allow Web Surfers To Make Random Jumps

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

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

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

— Perron says \( \exists \; \pi^T \geq 0 \) s.t. \( \pi^T = \pi^T \mathbf{S} \) 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

\[ S = \begin{pmatrix}
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/3 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix} \]

- 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 convergence

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 converge

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

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

\[ G = \alpha H + uv^T > 0 \]

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

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

Allow A Random Jump From Any Page

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

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

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

Some Happy Accidents

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

Allow A Random Jump From Any Page

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

\[ G = \alpha H + uv^T > 0 \]

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

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

Some Happy Accidents

\[ x^T G = \alpha x^T H + \beta v^T \]

Sparse computations with the original link structure

\[ \lambda_2(G) = \alpha \]

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

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

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

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

Some Happy Accidents

- \( x^T G = \alpha x^T H + \beta v^T \) Sparse computations with the original link structure

- \( \lambda_2(G) = \alpha \) 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
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.

(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.

(On Page C1)

- Ahold’s problems deepened as U.S. authorities opened inquiries into accounting at the Dutch company’s U.S. foodservice unit.

- Fleming said the SEC upgraded to a formal investigation an inquiry into the food wholesaler’s trade practices with suppliers.

(On Page A2)

- Consumer confidence fell to its lowest level since 1993, hurt by energy costs, the terrorism threat and a stagnant job market.

(On Page A3)

- The industrials rebounded on Wall Street.

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 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 weapon. 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 1 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 and units in a show of force as U.S. officials traveled into Iraq’s north for an opposition conference.

- Powell said North Korea hasn’t started 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 A14)

- The FBI came under withering bipartisan criticism in a Senate Judiciary report in which Sen. Specter accused the agency of mishandling a bomb case.

Web Master

As the Web Spreads...

Total Internet users, by household, in millions

<table>
<thead>
<tr>
<th>Year</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>60.8</td>
<td>80.3</td>
<td>130.0</td>
<td>250.0</td>
<td>360.0</td>
<td>430.0</td>
</tr>
</tbody>
</table>

Sources: Forrester Research; Nielsen/NetRatings

Google’s U.S. Presence Expands

Top search engines, in millions of unique visitors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>39.4</td>
<td>38.8</td>
<td>38.8</td>
<td>36.8</td>
<td>38.6</td>
</tr>
<tr>
<td>Yahoo</td>
<td>13.3</td>
<td>22.0</td>
<td>27.0</td>
<td>30.0</td>
<td>30.0</td>
</tr>
<tr>
<td>MSN Search</td>
<td>6.4</td>
<td>8.6</td>
<td>9.6</td>
<td>10.6</td>
<td>11.6</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>2.5</td>
<td>3.0</td>
<td>3.5</td>
<td>4.0</td>
<td>4.5</td>
</tr>
<tr>
<td>Overture</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
<td>2.5</td>
<td>3.0</td>
</tr>
<tr>
<td>CNET</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Category: Top shopping-referral sites, in millions of referrals

<table>
<thead>
<tr>
<th>Category</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>12.6</td>
<td>13.0</td>
<td>14.0</td>
<td>15.0</td>
<td>16.0</td>
</tr>
<tr>
<td>Dealline</td>
<td>2.5</td>
<td>2.0</td>
<td>1.5</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>BizRate</td>
<td>1.3</td>
<td>1.2</td>
<td>1.1</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>Overture</td>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
<td>1.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Epinions</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>1.1</td>
</tr>
<tr>
<td>CNET</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Exoticleatherwear Gets Cut Off

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 technique of paid-for placement—a method widely criticized as misleading—had just become legal under a new law passed in California.

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’

By Michael Totty
And Mylene Mangalindan

WASHINGTON—The Bush administration is preparing supplemental spending requests totaling as much as $95 billion for a war on 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 review proposed budget requests with

By Greg Jaffe
And John D. McKinnon
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 through a series of links 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 bottom line.

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 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 the system. "When people try to take scoring into their own hands, that turns into a worse experience for users," says Matt Cutts, a Google software engineer.

Coding Trickery

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

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 fooled 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 but was quickly detected by faster search engines.

Mr. Brin, 29, one of Google's two founders and now its president of technology, appeared at a conference in San Francisco last summer to explain the importance of appearing near the top of search-engine results, especially on Google. She boned up on search techniques, visiting online discussion groups dedicated to search engines and reading what material she could find on the Web.

At first, Ms. Holman limited herself to modest changes, such as loading her page with hidden metatag coding that would help steer a search toward her site when a user entered words such as "halter tops" or "leather wrap." Since it didn't give her site a good ranking on Google, she didn't think much of it. But then, her site was penalized in determining its rankings, the efforts had little effect on her search results.

The next she received an e-mail advertisement from AutomatedLinks.com, a site that promised to send traffic "through the roof" by linking to more than 2,000 Web sites to hers. Aside from attracting customers, the links were designed to improve her search results.

"SearchKing," as the company is known, "is a great way to get on an ad network and boost your rankings in Google and other search results," says Mr. Massa, SearchKing's chief executive, last August. "It's a great tool for PageRank, which is used by other search engines on a scale of one to 10 based on their popularity, and the rankings can be viewed by Web users if they install special Google software. PR Ad Network sells ads that are priced according to a site's PageRank, with higher-ranked sites commanding higher prices. When a site buys an advertising link on a highly ranked site, the ad buyer could see its ratings improve because of the greater weight Google gives to that link.

The high-stakes fight between Google and the optimizers can leave some Web-site owners confused. "I don't know how people are supposed to judge what is right and wrong," says Mr. McLean.

AutomatedLinks didn't respond to requests for comment. Google, Cutts, and the court, holding its own, said they didn't know about the case.

Mr. Cutts, the Google engineer, warns that the rules are clear and that it's better to follow them rather than try to get a problem fixed after a site has been penalized. "We want to return the most relevant pages we can," Mr. Cutts said. "The best you can do for a site owner is to do that following our guidelines.

Crackdown

Google has been stepping up its enforcement since 2001. It warned 19,000 sites that it had gotten its sites kicked out of the Google index and it provided a list of forbidden activities, including hiding text and "link schemes," such as the link farms. Google also warned against "cloaking," the practice of showing a different page to a visitor's computer than it's designed to score well while giving visitors a different, more attractive page—or creating multiple Web addresses that take visitors to a single site.

Mr. Agnell, the 58-year-old Fiat patriarch Gianni Agnell last month, was widely expected to take over as chairman, later this year. Mr. Agnell has served as chairmen...
Personalization is Coming

The Wall Street Journal
Search Engines Seek to Get Inside Your Head

April 25, 2007

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

By JESSICA E. VASCELLARO
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...
Conclusion

Google Augments PR With Content Scores For Final Rankings

Content “Metrics” Are Proprietary — But Known Examples

- Whether query terms appear in the title or the body
- Number of times query terms appear in a page
- Proximity of multiple query words to one another
- Appearance of query terms in a page (e.g., headings in bold font score higher)
- Content of neighboring web pages

Elegant and Exciting Application of Linear Algebra

That Is Changing The World