Google's PageRank
and Beyond

Carl Meyer

Department of Mathematics
North Carolina State University
Raleigh, NC

Colorado State University
Fort Collins, Colorado
November 2, 2007

Beautiful mathematics eventually tends to be useful.

And useful mathematics eventually tends to be beautiful.

Eye Of The Beholder

$$F(f) = \int_{-\infty}^{\infty} x(t) \mathrm{e}^{-\mathrm{i}2\pi f t} dt \quad \longleftrightarrow$$

Eye Of The Beholder

{ Perron–Frobenius
  Markov Chains } $\longleftrightarrow$ 

System for the Mechanical Analysis and Retrieval of Text

Harvard 1962 – 1965

Cornell 1965 – 1970

Gerard Salton

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

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} = # times term i appears in document j

Term–Document Matrix

$$\begin{array}{c} \\ \text{Term 1} \\ \text{Term 2} \\ \vdots \\ \text{Term m} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Doc 1} & \text{Doc 2} & \cdots & \text{Doc n} \end{array} \\ \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{pmatrix} \end{array} = \mathbf{A}_{m \times n}$$

Query Matching

Query Vector

$$\mathbf{q}^T = (q_1, q_2, \ldots, q_m) \qquad q_i = \begin{cases} 1 & \text{if Term } i \text{ is requested} \\ 0 & \text{if not} \end{cases}$$

Query Matching

Query Vector

$$\mathbf{q}^T = (q_1, q_2, \ldots, q_m) \qquad q_i = \begin{cases} 1 & \text{if Term } i \text{ is requested} \\ 0 & \text{if not} \end{cases}$$

How Close is Query to Each Document?

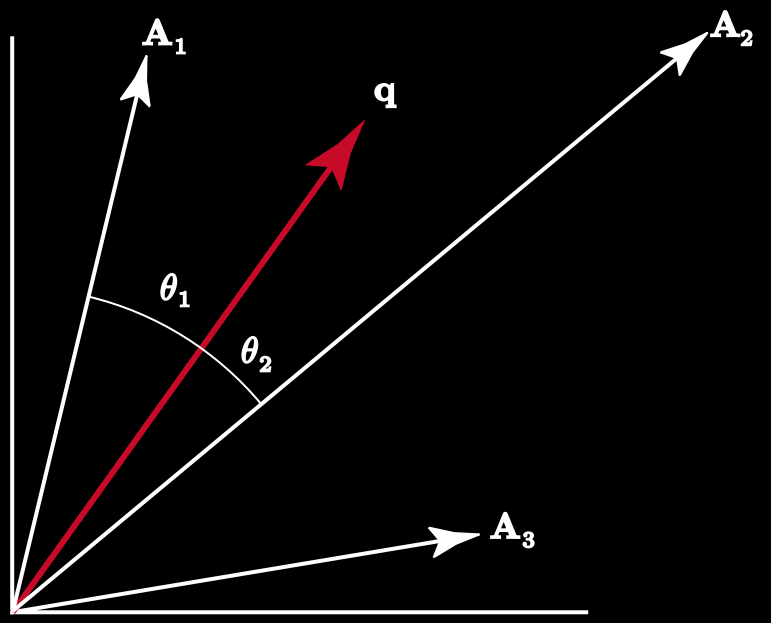
Query Matching

Query Vector

$$\mathbf{q}^T = (q_1, q_2, \ldots, q_m) \qquad q_i = \begin{cases} 1 & \text{if Term } i \text{ is requested} \\ 0 & \text{if not} \end{cases}$$

How Close is Query to Each Document?

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



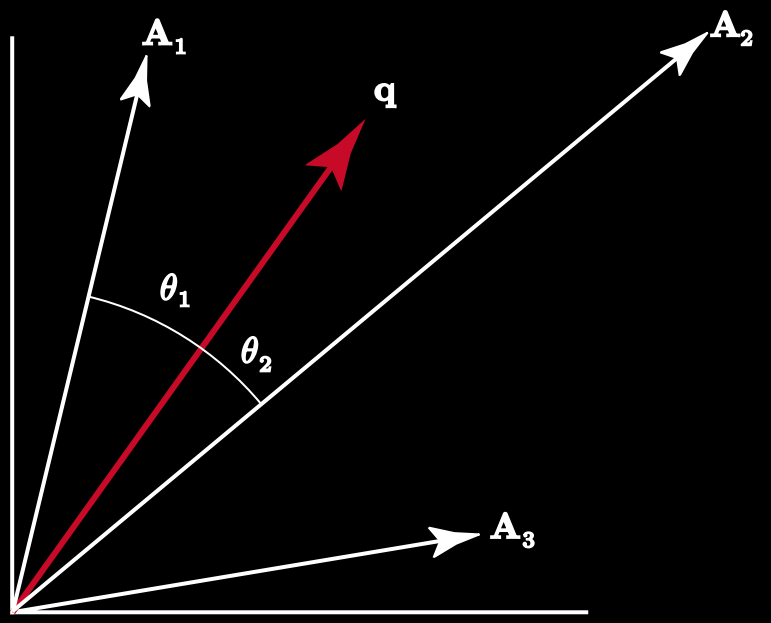
Query Matching

Query Vector

$$\mathbf{q}^T = (q_1, q_2, \ldots, q_m) \qquad q_i = \begin{cases} 1 & \text{if Term } i \text{ is requested} \\ 0 & \text{if not} \end{cases}$$

How Close is Query to Each Document?

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



Use $\delta_i = \cos\theta_i = \dfrac{\mathbf{q}^T\mathbf{A}_i}{\|\mathbf{q}\|\,\|\mathbf{A}_i\|}$

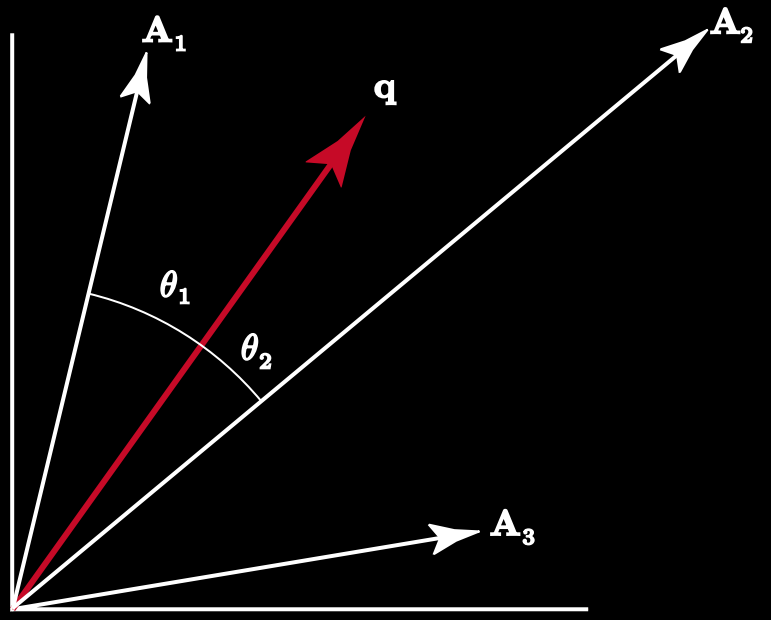
Query Matching

Query Vector

$$\mathbf{q}^T = (q_1, q_2, \ldots, q_m) \qquad q_i = \begin{cases} 1 & \text{if Term } i \text{ is requested} \\ 0 & \text{if not} \end{cases}$$

How Close is Query to Each Document?

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



Use $\delta_i = \cos\theta_i = \dfrac{\mathbf{q}^T\mathbf{A}_i}{\|\mathbf{q}\|\,\|\mathbf{A}_i\|}$

Rank documents by size of δ_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
  - Computer information retrieval using latent semantic structure. U.S. Patent No. 4,839,853, June 13, 1989.
  - Computerized cross-language document retrieval using latent semantic indexing. U.S. Patent No. 5,301,109, April 5, 1994.
- LATENT SEMANTIC INDEXING

Latent Semantic Indexing

Use a finite Fourier expansion of A

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{Z}_i, \qquad \langle \mathbf{Z}_i | \mathbf{Z}_j \rangle = \begin{cases} 1 & i=j, \\ 0 & i \neq j, \end{cases} \qquad |\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_r|$$

$|\sigma_i| = |\langle \mathbf{Z}_i | \mathbf{A} \rangle|$ = amount of **A** in direction of $\mathbf{Z}_i$

Latent Semantic Indexing

Use a finite Fourier expansion of A

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{Z}_i, \qquad \langle \mathbf{Z}_i | \mathbf{Z}_j \rangle = \begin{cases} 1 & i=j, \\ 0 & i \neq j, \end{cases} \qquad |\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_r|$$

$$|\sigma_i| = |\langle \mathbf{Z}_i | \mathbf{A} \rangle| = \text{amount of } \mathbf{A} \text{ in direction of } \mathbf{Z}_i$$

Realign data along dominant directions $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_k, \mathbf{Z}_{k+1}, \ldots, \mathbf{Z}_r\}$

- Project **A** onto $span\,\{\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k\}$

Latent Semantic Indexing

Use a finite Fourier expansion of A

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{Z}_i, \qquad \langle \mathbf{Z}_i | \mathbf{Z}_j \rangle = \begin{cases} 1 & i=j, \\ 0 & i \neq j, \end{cases} \qquad |\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_r|$$

$|\sigma_i| = |\langle \mathbf{Z}_i | \mathbf{A} \rangle|$ = amount of **A** in direction of $\mathbf{Z}_i$

Realign data along dominant directions $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_k, \mathbf{Z}_{k+1}, \ldots, \mathbf{Z}_r\}$

— Project **A** onto $span\{\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k\}$

Truncate: $\mathbf{A}_k = P(\mathbf{A}) = \sigma_1 \mathbf{Z}_1 + \sigma_2 \mathbf{Z}_2 + \cdots + \sigma_k \mathbf{Z}_k$

Latent Semantic Indexing

Use a finite Fourier expansion of A

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{Z}_i, \qquad \langle \mathbf{Z}_i | \mathbf{Z}_j \rangle = \begin{cases} 1 & i=j, \\ 0 & i \neq j, \end{cases} \qquad |\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_r|$$

$$|\sigma_i| = |\langle \mathbf{Z}_i | \mathbf{A} \rangle| = \text{amount of } \mathbf{A} \text{ in direction of } \mathbf{Z}_i$$

Realign data along dominant directions $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_k, \mathbf{Z}_{k+1}, \ldots, \mathbf{Z}_r\}$

— Project **A** onto $span\{\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k\}$

Truncate: $\mathbf{A}_k = P(\mathbf{A}) = \sigma_1 \mathbf{Z}_1 + \sigma_2 \mathbf{Z}_2 + \cdots + \sigma_k \mathbf{Z}_k$

LSI: Query matching with $\mathbf{A}_k$ in place of A

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

Latent Semantic Indexing

Use a finite Fourier expansion of A

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{Z}_i, \qquad \langle \mathbf{Z}_i | \mathbf{Z}_j \rangle = \begin{cases} 1 & i=j, \\ 0 & i \neq j, \end{cases} \qquad |\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_r|$$

$|\sigma_i| = |\langle \mathbf{Z}_i | \mathbf{A} \rangle|$ = amount of **A** in direction of $\mathbf{Z}_i$

Realign data along dominant directions $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_k, \mathbf{Z}_{k+1}, \ldots, \mathbf{Z}_r\}$

— Project **A** onto $span\{\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k\}$

Truncate: $\mathbf{A}_k = P(\mathbf{A}) = \sigma_1 \mathbf{Z}_1 + \sigma_2 \mathbf{Z}_2 + \cdots + \sigma_k \mathbf{Z}_k$

LSI: Query matching with $\mathbf{A}_k$ in place of A

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

"Best" mathematical solution

— SVD: $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ $\qquad\qquad$ $\mathbf{Z}_i = \mathbf{u}_i \mathbf{v}_i^T$

Strengths & Weaknesses

Pros

- Finds hidden connections

Strengths & Weaknesses

Pros

- Finds hidden connections
- Can be adapted to identify document clusters
  - Data mining applications

Strengths & Weaknesses

Pros

- Finds hidden connections
- Can be adapted to identify document clusters
  - Data 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
  - Data 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
  - Data 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
  - Susceptible to malicious manipulation

Strengths & Weaknesses

Pros

- Finds hidden connections
- Can be adapted to identify document clusters
  - — Data 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
  - — Susceptible to malicious manipulation
- Difficult to add & delete documents

Strengths & Weaknesses

Pros

- Finds hidden connections
- Can be adapted to identify document clusters
  - — Data 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
  - — Susceptible to malicious manipulation
- Difficult to add & delete documents
- Finding optimal compression requires empirical tuning

Web Facts

Different from other document collections

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

Web Facts

Different from other document collections

- It's huge
  - Over 10 billion pages, where average page size $\approx$ 500KB
  - 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 Facts

Different from other document collections

- It's huge
  - Over 10 billion pages, where average page size $\approx$ 500KB
  - 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 Facts

Different from other document collections

- It's huge
  - Over 10 billion pages, where average page size $\approx$ 500KB
  - 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!
- 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

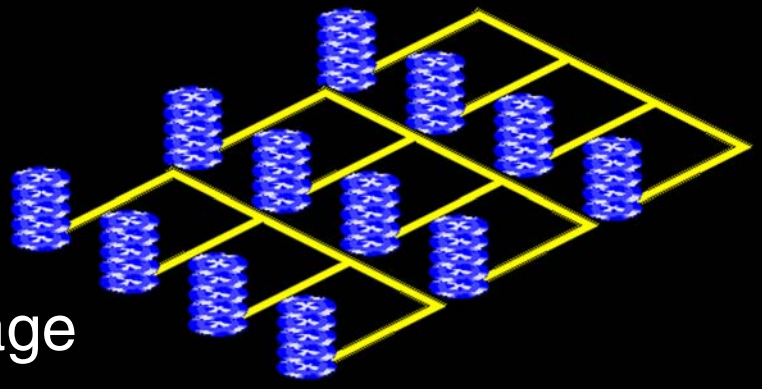
Doc Server

Stores docs
and snippits

Index Server

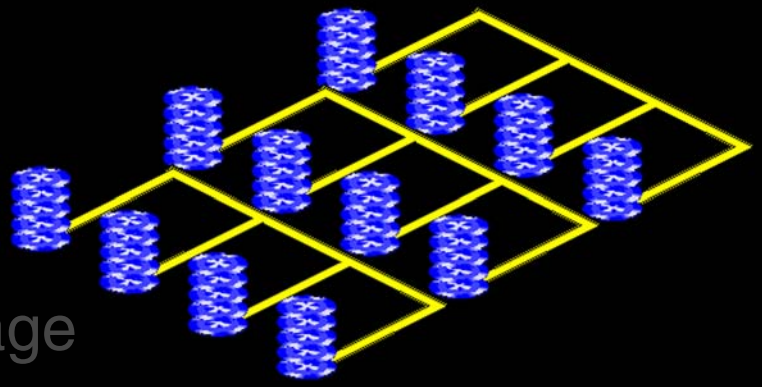
Scans pages and does term indexing
Terms $\longrightarrow$ Pages (similar to book index)

The Ranking Module



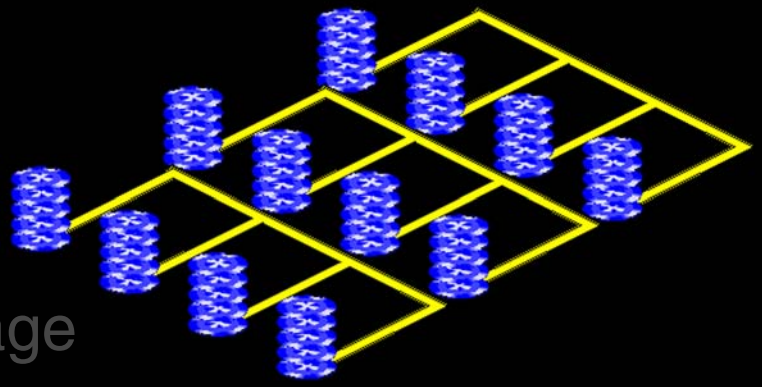
- Measure the importance of each page

The Ranking Module



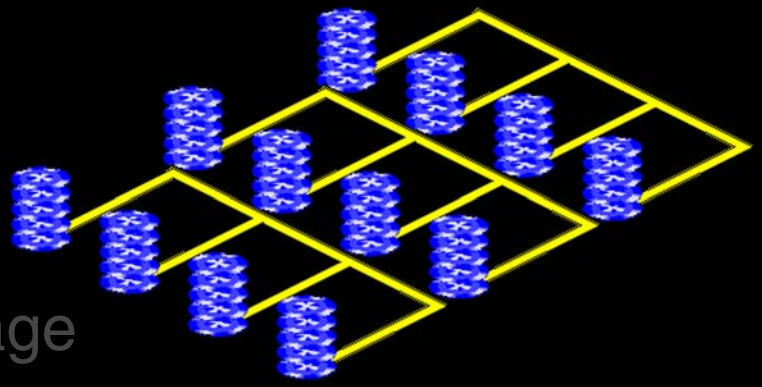
- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations

The Ranking Module



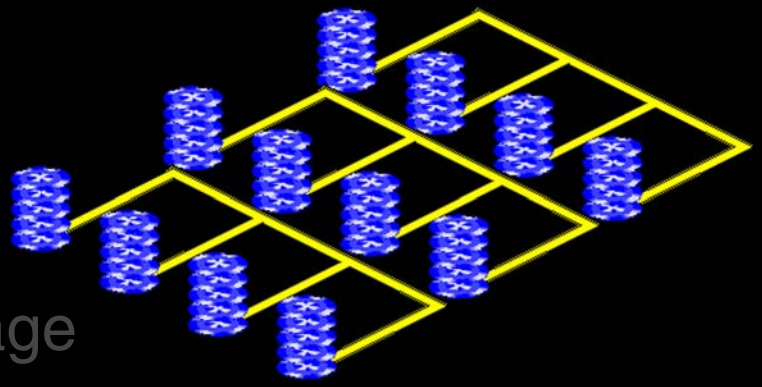
- Measure the importance of each page
- The measure should be Independent of any query
  - — Primarily determined by the link structure of the Web
  - — Tempered by some content considerations
- Compute these measures off-line long before any queries are processed

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

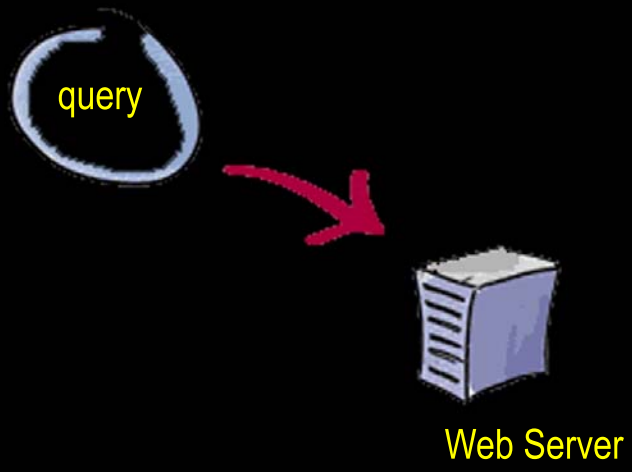
The Ranking Module



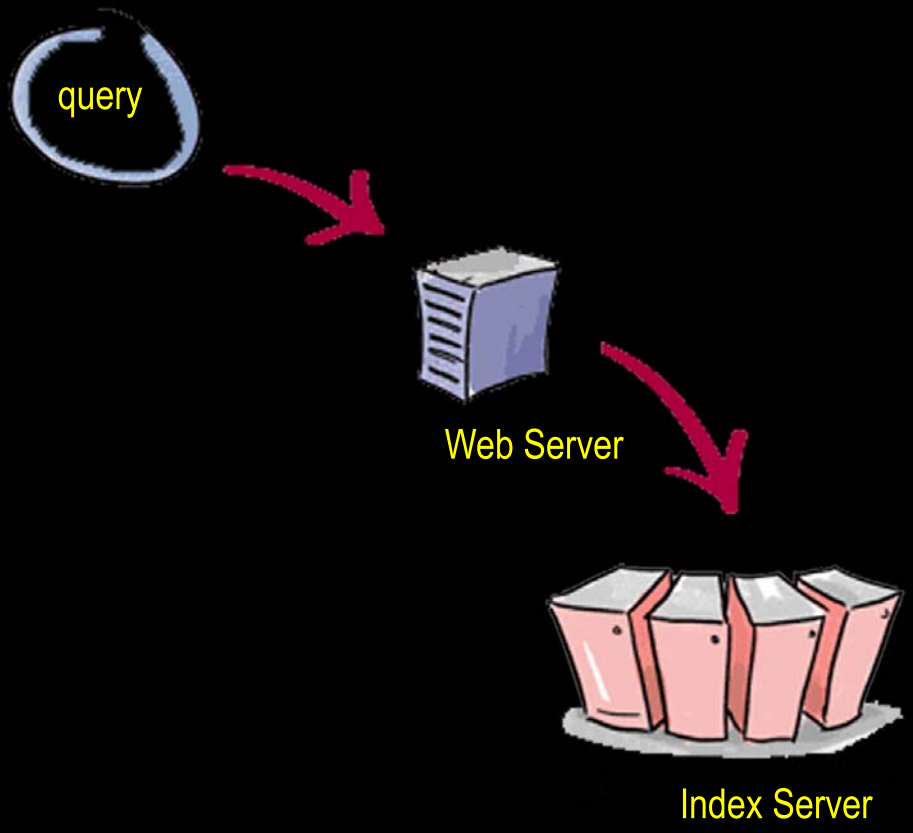
- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
- Compute these measures off-line long before any queries are processed
- Google's PageRank© technology distinguishes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

Google's PageRank = Google's $$$$$

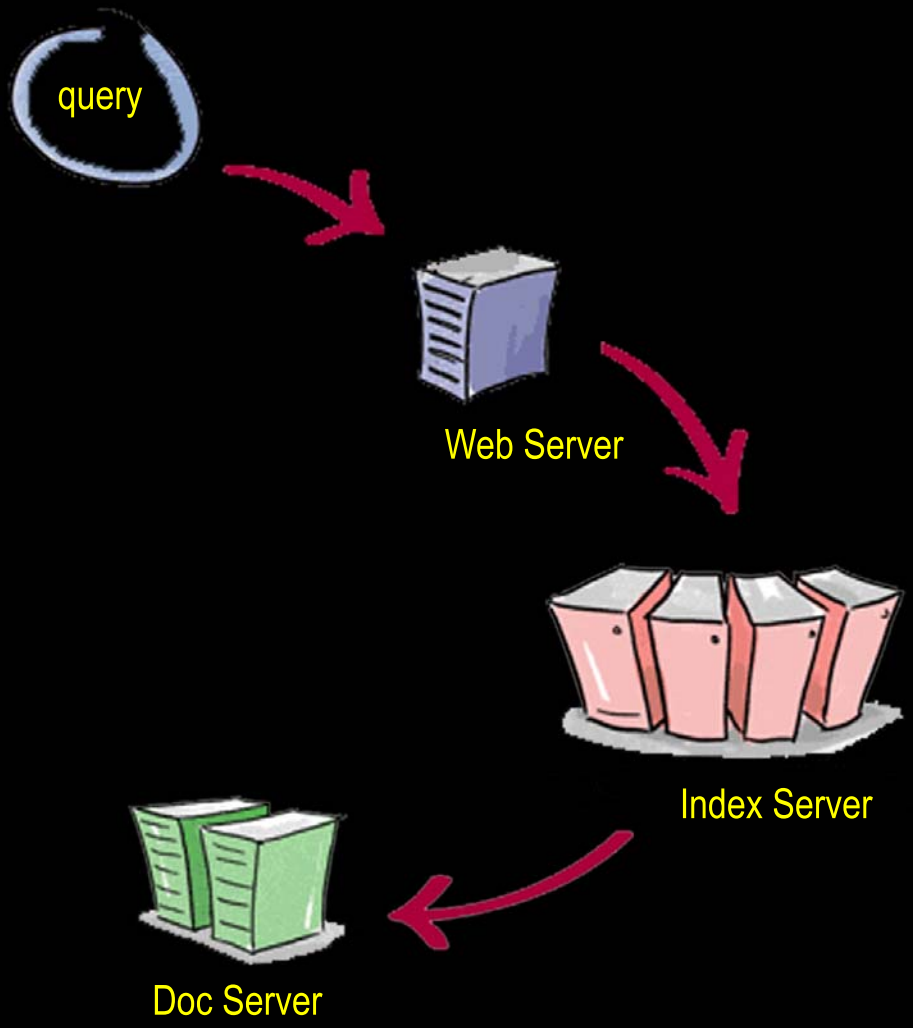
The Process



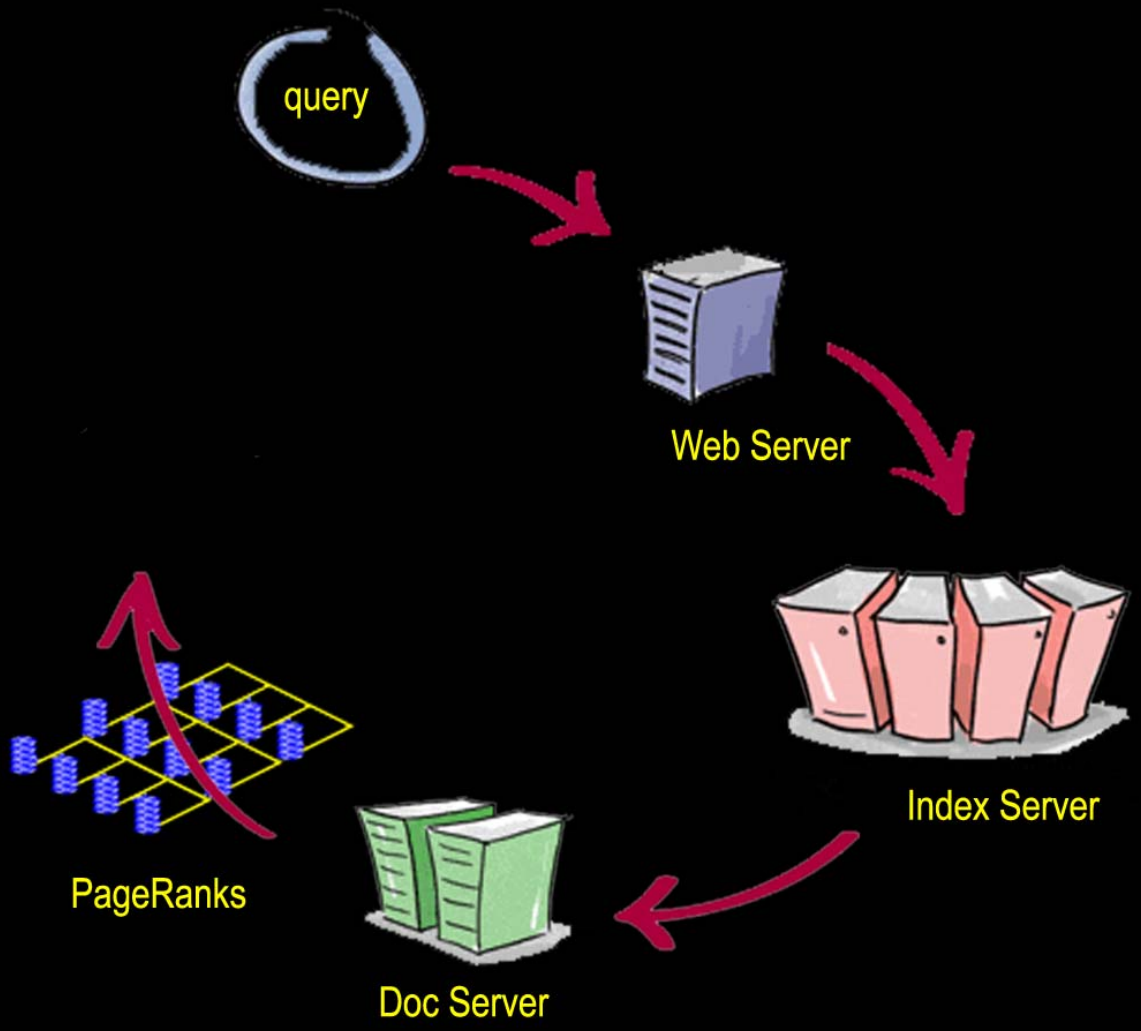
The Process



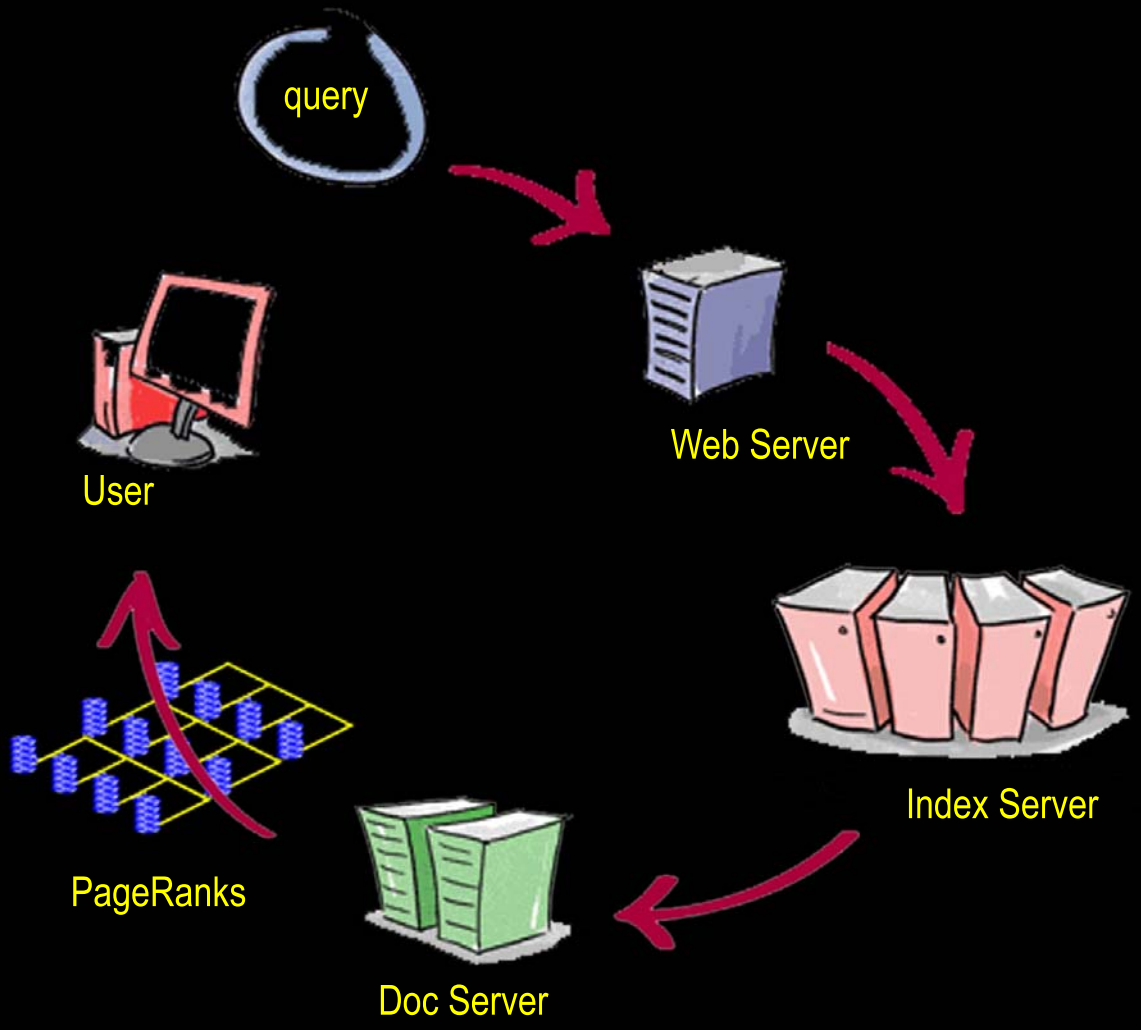
The Process

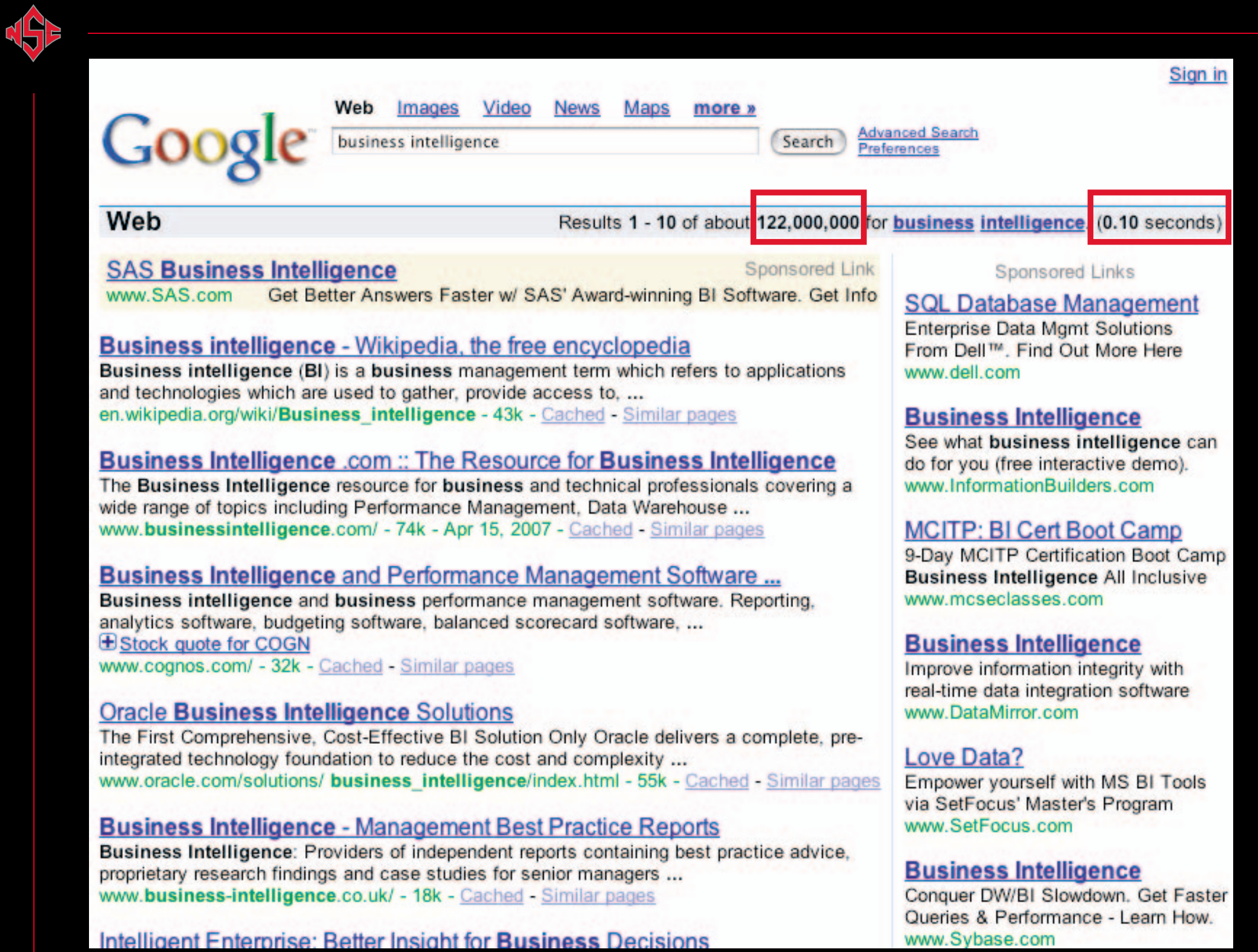


The Process



The Process

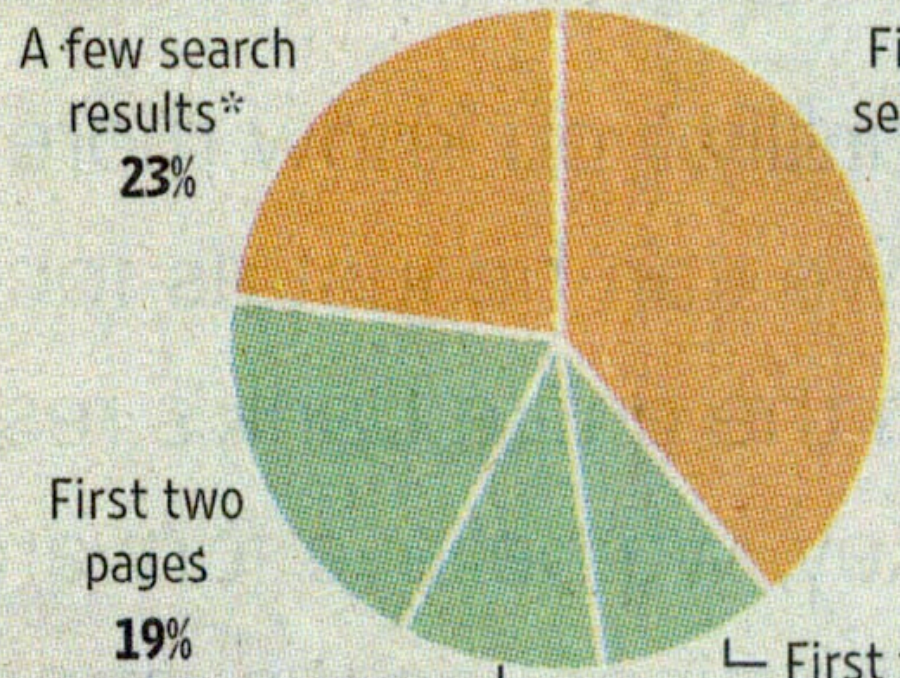



Sign in
Google
Web Images Video News Maps more »
business intelligence
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)
SAS Business Intelligence
Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info
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
Sponsored Links
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
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com

The Wall Street Journal, April 13, 2007

Take Your Pick

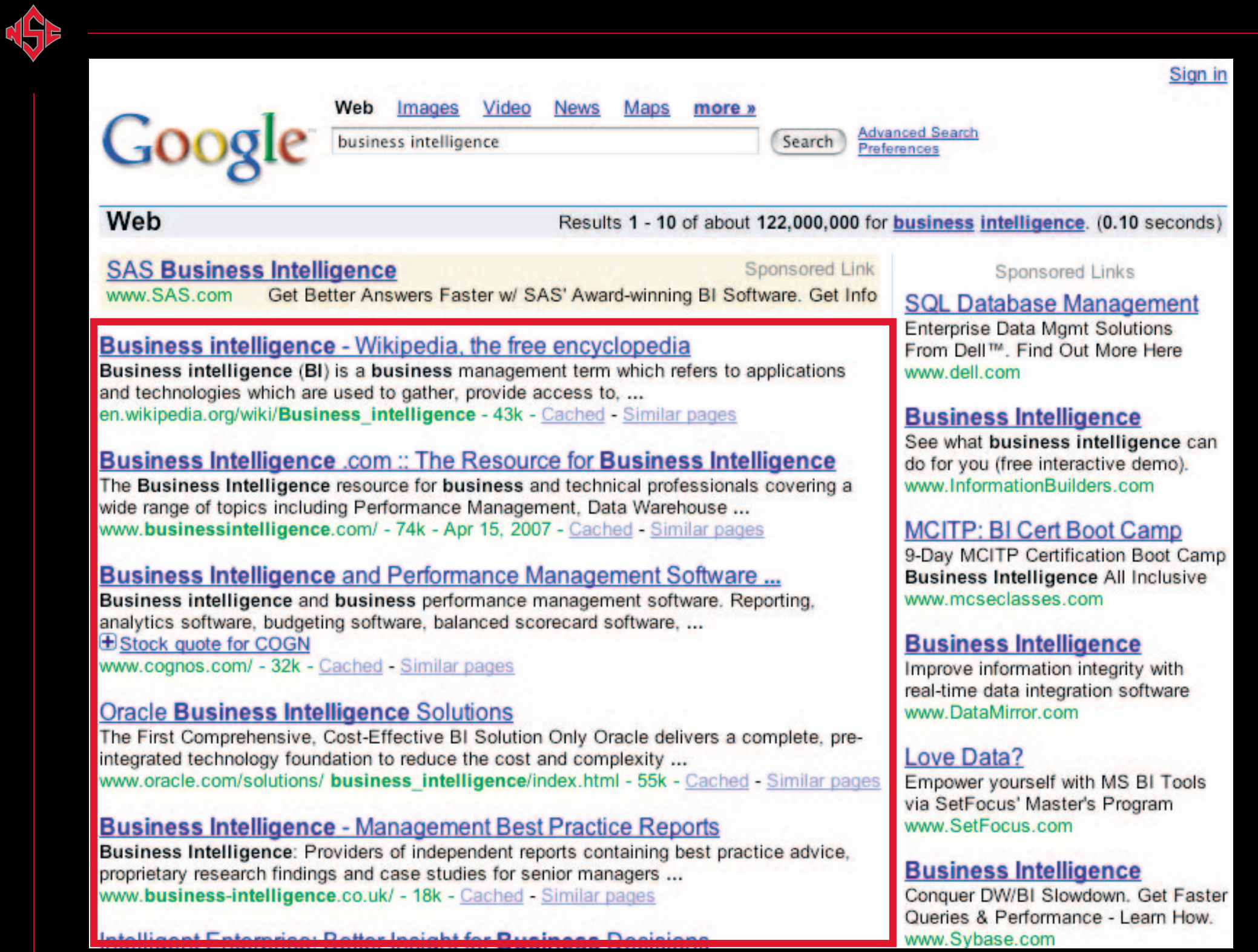
Amount of Internet search results that Web surfers typically scan before selecting one



*Top results without reading through the whole page

Note: Sample size is 2,369 people

Sources: JupiterResearch; iProspect



Sign in

Google

Web Images Video News Maps more »

business intelligence

Search Advanced Search Preferences

Web

Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)

Sponsored Link

SAS Business Intelligence

www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info

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

Sponsored Links

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

Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.

www.Sybase.com

Sign in

Google

Web Images Video News Maps more »

business intelligence Search Advanced Search Preferences

Web

Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)

Sponsored Link

SAS Business Intelligence
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info

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

Sponsored Links

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
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com

Sign in

Google

Web Images Video News Maps more »

business intelligence Search Advanced Search Preferences

Web Results **1 - 10** of about **122,000,000** for **business intelligence**. (**0.10** seconds)

Sponsored Link

SAS Business Intelligence
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info

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

Sponsored Links

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

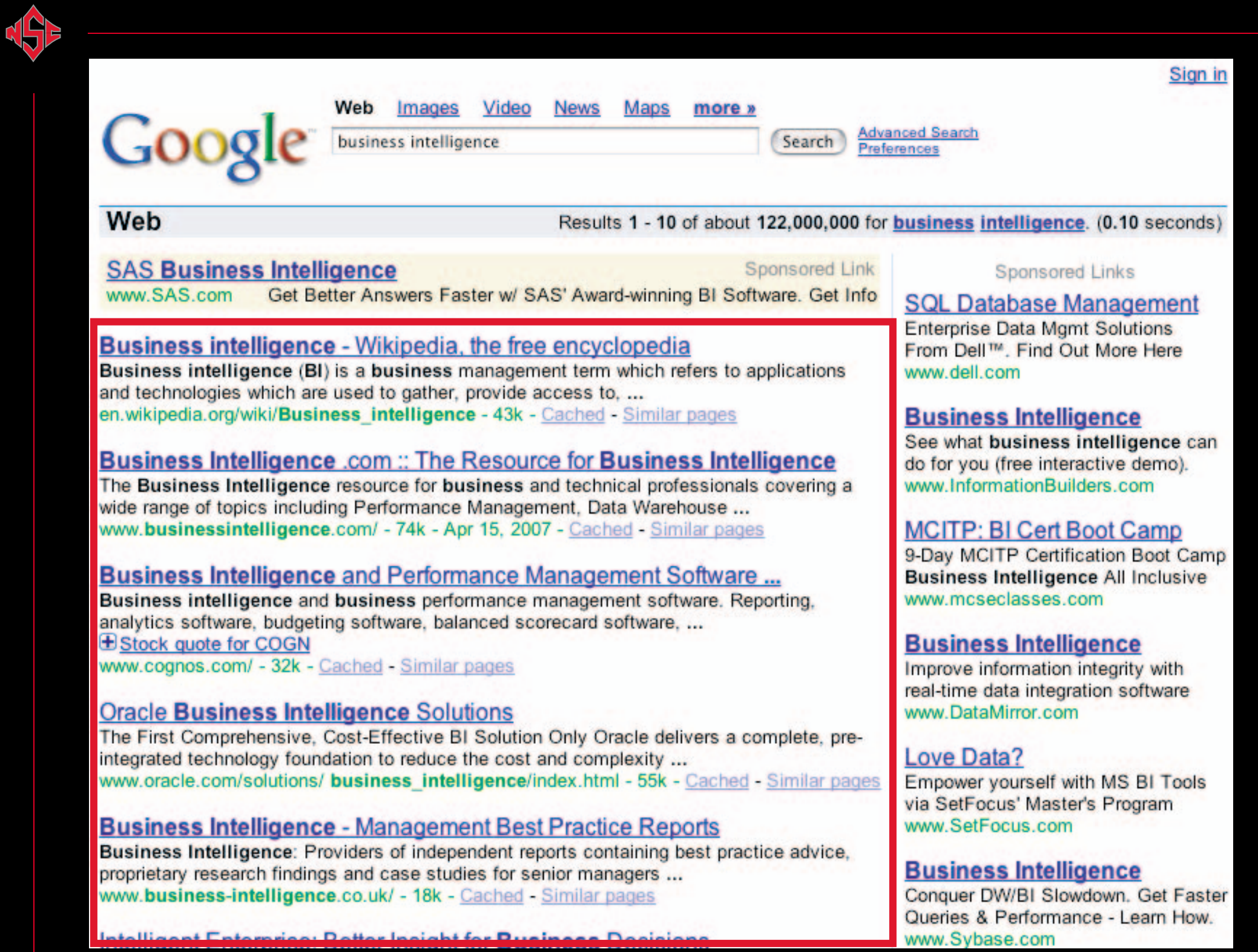
Business Intelligence
See what **business intelligence** can do for you (free interactive demo).
www.InformationBuilders.com

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
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com



Sign in

Google

Web Images Video News Maps more »

business intelligence

Search Advanced Search Preferences

Web

Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)

Sponsored Link

SAS Business Intelligence

www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info

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

Sponsored Links

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

Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.

www.Sybase.com

LEADING THE NEWS

Google Profit Back in Fast Lane

Core Search, Ad Brokering Fuel 46% Increase in Net Amid Sharp Hiring Pace

By Kevin J. Delaney

Google Inc. shook off its previous quarter's earnings woes and increased its profit-growth rate, even while accelerating its rapid hiring pace.

The Mountain View, Calif., Internet company said third-quarter profit rose 46%, compared with 28% profit growth in the previous quarter, thanks to the continued strength in its core search-advertising business and its brokering of ads that appear on partner Web sites. Revenue jumped 57%.

"We're very pleased with such strong results in what's seasonally one of our weaker quarters," Chief Executive Eric Schmidt said in a conference call with analysts. "Looking back at the quarter, it's obvious to us our model continues to work very well." He and other Google executives highlighted the company's increased focus on selling additional types of ads, such as those appearing on its YouTube video-sharing service and television commercials that appear through satellite-TV provider **EchoStar Communications** Corp.

"The quarter was strong and clean," said Youssef Squali, an Internet analyst at Jefferies & Co. in New York, whose firm makes a market in Google shares. "Looks like this well-oiled machine will continue to impress for some time to come."

The earnings report sent shares of Google, which reported results after the close of regular trading, up 1.2% to $647 in after-hours trading. Google traded at $639.62 in 4 p.m. Nasdaq Stock Market composite trading. Its shares have been on a tear since September and have risen almost 40% this year.

Google added 2,130 employees during the quarter to end at 15,916, compared with its addition of 1,548 employees during the second quarter. The company's second-quarter profit had fallen short of expectations as costs, including staff expenses, weighed on the bottom line, partly because Google hired faster than it had expected.

"The numbers you're seeing are essentially an overhang...from hiring that had been agreed to many, many months earlier," said Mr. Schmidt. "Going forward you should feel comfortable we are paying a lot of attention to the headcount."

"Given what the revenue did I would have thought more [of it] would've dropped down to the bottom line," said Doug Anmuth, an analyst with Lehman Brothers Inc., whose firm has received non-investment banking compensation from Google and makes a market in its shares. Mr. Anmuth said it was "a pretty solid quarter" but profits appeared to have been reduced by the significant staffing additions.

Excluding certain stock-based compensation and other factors, Google earned $3.91 a share, above a Wall Street average forecast of $3.78 on that basis, according to Thomson Financial. Revenue rose to $4.23 billion from $2.69 billion a year earlier.

Google's revenue excluding commissions paid to marketing partners totaled $3.01 billion, beating analysts' estimates of $2.94 billion. The company's rate of revenue growth roughly stabilized in the third quarter. The increase of 57% from a year earlier compares with 58% in the second quarter and 73% for all of 2006. The rate of revenue growth from partner sites that carry advertising brokered by Google picked up during the quarter to 40% from 36% during the second quarter.

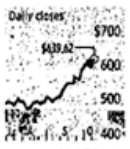
International operations accounted for 48% of revenue during the third quarter, in line with the second quarter.

Google handled 53.6% of U.S. Web-search queries in August, compared with 52.7% in June, NetRatings Inc. said.

Google

Net: ▲ 46% (GOOG)

	3Q'07	3Q'06
Net:†	1,070	733
EPS ($):	3.38	2.36
Est ($):‡	3.78	
Rev:†	4,231	2,690
Divisional breakdown:†		
Advertising	4,189	2,663
Licensing	42	27



† In millions of dollars. ‡Analysts' consensus estimate provided by Thomson Financial; may exclude certain items or discontinued operations

LEADING THE NEWS

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 brokering 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 quarter from their year-earlier periods. Revenue had risen 34% during the first quarter of 2006 from the year before.

J.P. Morgan Securities Inc. analyst Imran Khan said that the growth in revenue from graphical display ads such as banners is slowing for Yahoo amid competition from rival Internet companies. However, "you have to expect Panama will offset some of the weakness you will see on the graphical advertising side in the coming quarters," said Mr. Khan, whose company has an investment banking relationship with Yahoo.

"We are very pleased with the initial progress of Panama," said Chief Financial Officer Susan Decker in a conference call with analysts, adding that Yahoo's average revenue for each search it handled had begun improving by the end of the first quarter. She said that Yahoo expects to see a moderate decline in the average revenue it generates from graphical display advertising for each Web page viewed by a user. The company predicts the revenue growth rate overall will stem its decline and resume rising starting in the second quarter, thanks to the search-advertising improvements.

Yahoo

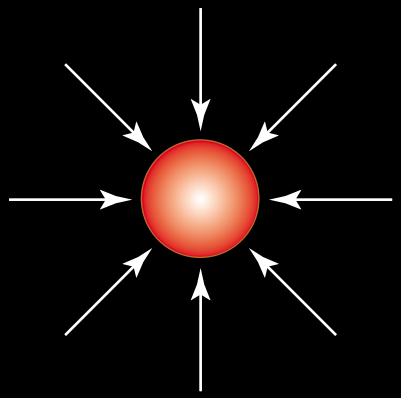
Net: ▼ 11% (YHOO)

	1Q'07	1Q'06
Net:†	142	160
EPS ($):	0.10	0.11
Est ($):‡	0.11	
Rev:†	1,672	1,567
Divisional breakdown:†		
U.S.	1,469	1,381
Int'l	203	186

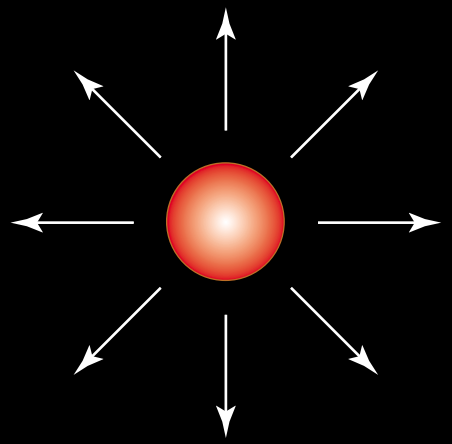
† In millions of dollars. ‡Analysts' consensus estimate provided by Thomson Financial; may exclude certain items or discontinued operations

How To Measure “Importance”

Landmark Result Paper

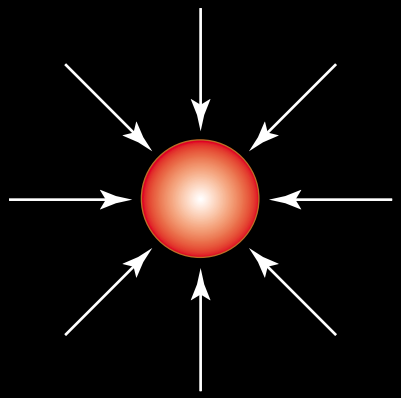


Survey Paper—Big Bib



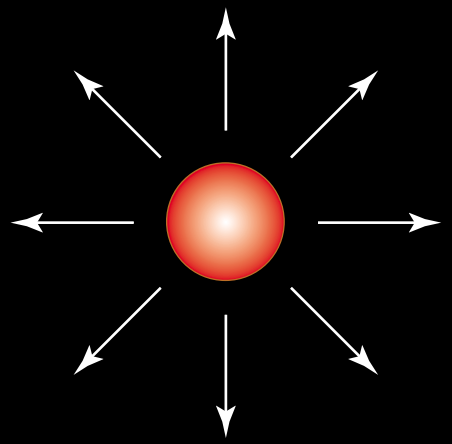
How To Measure “Importance”

Landmark Result Paper



Authorities

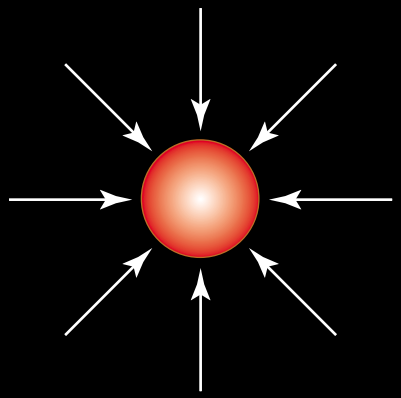
Survey Paper—Big Bib



Hubs

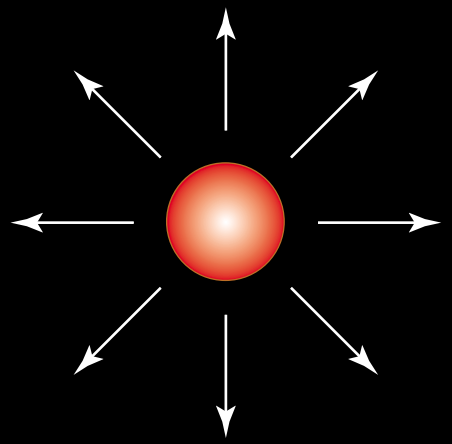
How To Measure “Importance”

Landmark Result Paper



Authorities

Survey Paper—Big Bib



Hubs

- Good hubs point to good authorities
- Good authorities are pointed to by good hubs

HITS

Hypertext Induced Topic Search (1998)

Jon Kleinberg

Determine Authority & Hub Scores

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

HITS

Hypertext Induced Topic Search (1998)

Jon Kleinberg

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 $P_i \quad \Rightarrow \quad \mathbf{h_0} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

HITS

Hypertext Induced Topic Search (1998)

Jon Kleinberg

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 $P_i \quad \Rightarrow \quad \mathbf{h_0} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
- Define Authority Scores (on the first pass)

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

HITS

Hypertext Induced Topic Search (1998)

Jon Kleinberg

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 $P_i \quad \Rightarrow \quad \mathbf{h_0} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
- Define Authority Scores (on the first pass)

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

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

HITS Algorithm

Refine Hub Scores

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

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

HITS Algorithm

Refine Hub Scores

- $h_i = \sum_{j:P_i \to P_j} a_j \quad \Rightarrow \quad \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

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

Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_1 = \mathbf{L}^T\mathbf{h}_0$

HITS Algorithm

Refine Hub Scores

- $h_i = \sum_{j:P_i \to P_j} a_j \quad \Rightarrow \quad \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

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

Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_1 = \mathbf{L}^T\mathbf{h}_0$
  - $\mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

HITS Algorithm

Refine Hub Scores

- $h_i = \sum_{j:P_i \to P_j} a_j \quad \Rightarrow \quad \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

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

Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_1 = \mathbf{L}^T\mathbf{h}_0$
  - $\mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$
    - $\mathbf{a}_2 = \mathbf{L}^T\mathbf{h}_1$

HITS Algorithm

Refine Hub Scores

- $h_i = \sum_{j:P_i \to P_j} a_j \quad \Rightarrow \quad \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$ $\qquad L_{ij} = \begin{cases} 1 & P_i \to P_j \\ 0 & P_i \not\to P_j \end{cases}$

Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_1 = \mathbf{L}^T\mathbf{h}_0$
  - $\mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$
    - $\mathbf{a}_2 = \mathbf{L}^T\mathbf{h}_1$
      - $\mathbf{h}_2 = \mathbf{L}\mathbf{a}_2$
        - ⋱

HITS Algorithm

Refine Hub Scores

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

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

Successively Re-refine Authority & Hub Scores

- $\mathbf{a_1 = L^T h_0}$
  - $\mathbf{h_1 = La_1}$
    - $\mathbf{a_2 = L^T h_1}$
      - $\mathbf{h_2 = La_2}$
      - ⋱

Combined Iterations

- $\mathbf{A = L^T L}$ (authority matrix)

HITS Algorithm

Refine Hub Scores

- $h_i = \sum_{j:P_i \to P_j} a_j \quad \Rightarrow \quad \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$ $\qquad L_{ij} = \begin{cases} 1 & P_i \to P_j \\ 0 & P_i \not\to P_j \end{cases}$

Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_1 = \mathbf{L}^T\mathbf{h}_0$
  - $\mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$
    - $\mathbf{a}_2 = \mathbf{L}^T\mathbf{h}_1$
      - $\mathbf{h}_2 = \mathbf{L}\mathbf{a}_2$
      - ⋱

Combined Iterations

- $\mathbf{A} = \mathbf{L}^T\mathbf{L}$ (authority matrix) $\qquad \mathbf{a}_k = \mathbf{A}\mathbf{a}_{k-1} \rightarrow$ e-vector (direction)

HITS Algorithm

Refine Hub Scores

- $h_i = \sum_{j:P_i \to P_j} a_j \quad \Rightarrow \quad \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$ $\qquad L_{ij} = \begin{cases} 1 & P_i \to P_j \\ 0 & P_i \not\to P_j \end{cases}$

Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_1 = \mathbf{L}^T\mathbf{h}_0$
  - $\mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$
    - $\mathbf{a}_2 = \mathbf{L}^T\mathbf{h}_1$
      - $\mathbf{h}_2 = \mathbf{L}\mathbf{a}_2$
      - ⋱

Combined Iterations

- $\mathbf{A} = \mathbf{L}^T\mathbf{L}$ (authority matrix) $\qquad \mathbf{a}_k = \mathbf{A}\mathbf{a}_{k-1} \to$ e-vector (direction)
- $\mathbf{H} = \mathbf{L}\mathbf{L}^T$ (hub matrix) $\qquad \mathbf{h}_k = \mathbf{H}\mathbf{h}_{k-1} \to$ e-vector (direction)

HITS Algorithm

Refine Hub Scores

- $h_i = \sum_{j:P_i \to P_j} a_j \quad \Rightarrow \quad \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$ $\qquad L_{ij} = \begin{cases} 1 & P_i \to P_j \\ 0 & P_i \not\to P_j \end{cases}$

Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_1 = \mathbf{L}^T\mathbf{h}_0$
  - $\mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$
    - $\mathbf{a}_2 = \mathbf{L}^T\mathbf{h}_1$
      - $\mathbf{h}_2 = \mathbf{L}\mathbf{a}_2$
        - $\ddots$

Combined Iterations

- $\mathbf{A} = \mathbf{L}^T\mathbf{L}$ (authority matrix) $\qquad \mathbf{a}_k = \mathbf{A}\mathbf{a}_{k-1} \to$ e-vector (direction)
- $\mathbf{H} = \mathbf{L}\mathbf{L}^T$ (hub matrix) $\qquad \mathbf{h}_k = \mathbf{H}\mathbf{h}_{k-1} \to$ e-vector (direction)

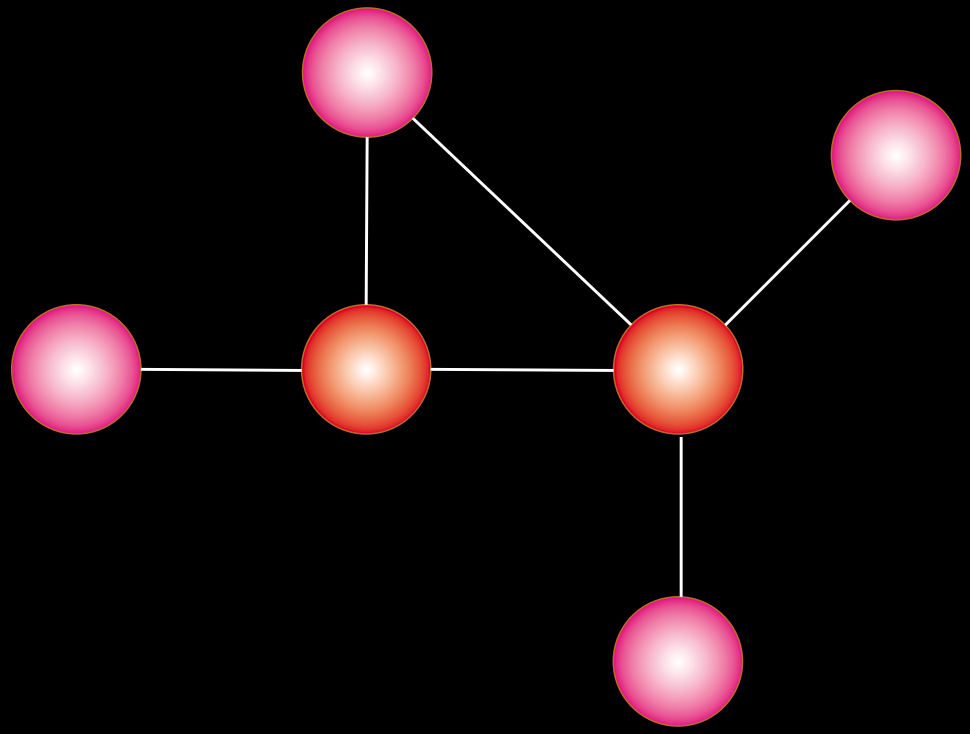
!! **May not be uniquely defined** !!

Compromise

1. Do direct query matching

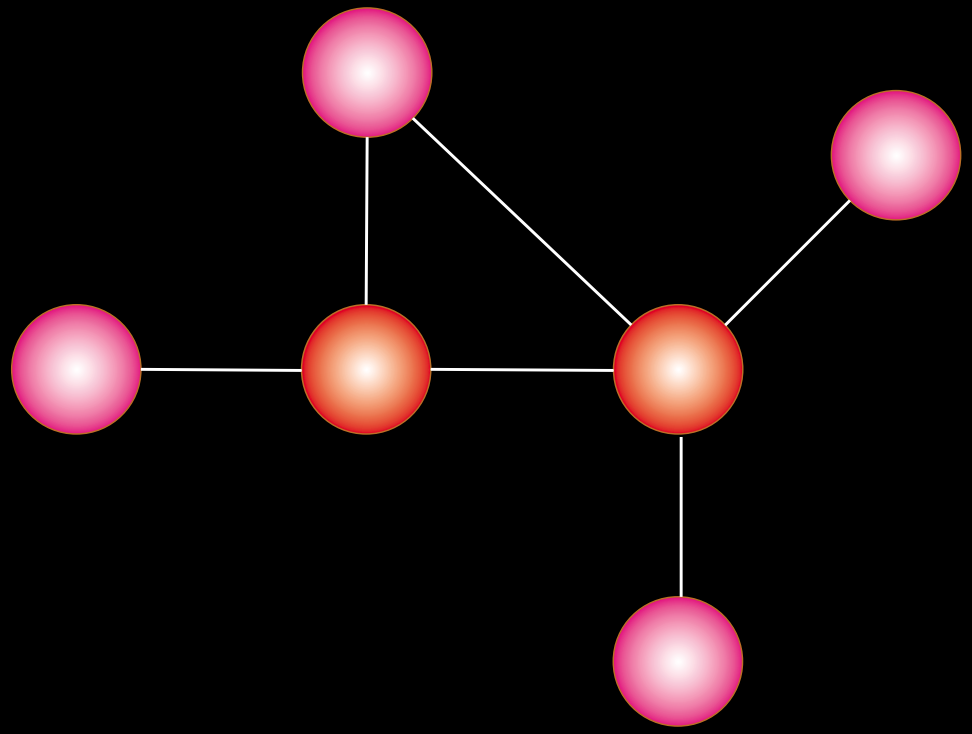
Compromise

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

Ask.com™

Every time you cough, a hunk of code or a piece of some obscure url comes shooting out. You can't see it, but it's there. Probably there is some on your shoes. A little string of binary code, or maybe the "r" and "g" from a dot org, right there on your burgundy cap-toes. The reason is that you're drowning in a sea of information. Heed not the worrisome findings of the recent ODP coastline study—by the time glacial melt brings the ocean to your doorstep, your lungs will already be full of html. $G=(v,e)$:

WE DON'T HAVE TO TELL YOU THE WORLD WIDE WEB IS AN ANARCHIC FORM OF POPULIST HYPERMEDIA.

But we WILL tell you it's a hypertext corpus of unfathomable intricacy, and it's expanding faster than a flat universe in a cosmologically significant vacuum energy density. For the love of Gödel, just look at the thing! Millions of participants with as many agendas, cranking out hyperlinked content like there's no tomorrow. In fact, at this rate, the disappearance of tomorrow, or at least a universally accepted definition thereof, is actually a valid concern.

SEARCH IS AN UNDERSTATEMENT. ODYSSEAN QUEST IS MORE LIKE IT.

So how are you supposed to find anything in this great roiling miasma of ones and zeros? Text-based searches are not so good. If you believe otherwise, consider the word facial. A search engine that takes nothing more than the word itself into account will return

$$A=\begin{bmatrix}0&0&1&1&0&0\\0&0&0&1&1&1\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix}$$

textually consistent but conceptually scattered results. On one end of the facial spectrum, there's a mud mask. The other kind of facial, well...as anyone who rolls sans adult filter can attest, it's a different deal altogether. Look, even if you do manage to cluster a word into five different meanings, there's still the fact that each individual meaning yields nearly infinite search results. And a quindecillion divided by five is still two hundred quattuordecillion.

ALL OF A SUDDEN, "WHO KNOWS?" IS AN ASTUTE QUESTION.

Searching the Internet, it turns out, is not much different from searching the real world. The best thing to do is ask someone who knows. An authority on the subject. But who are the authorities, and what qualifies them as such in the first place? A Web page can't just declare itself an authority. If authority could be generated endogenously, Louis de Branges would have verified his own proof of the Riemann Hypothesis. Neither should authority be conferred from one page to another. This means you'd be OK letting Herman Mudgett pick your primary care guy. Last in the triumvirate of really-bad-ways-to-determine-authority is the notion of popularity. Surprisingly, this is the method employed by today's most widely used search engines. They find sites with the most links and present them as authorities. This is roughly analogous to handing the Fields Medal to your high school homecoming queen.

THE ANSWER CAME FROM BOOKS. WEIRD.

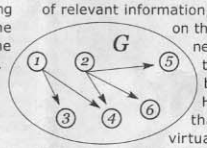
So what's the solution to search? While computer science was trying to coax an answer from its collective hard drive, it was sitting right there in the stacks all along. Who could have guessed that when Eugene Garfield went all bibliometric and devised a system to find out how much a journal mattered by counting the number of times that journal was cited in other publications, he consciously invented the beginnings of a system that might work in search. Then Gabriel Pinski and Francis Narin took it a step further by suggesting some citations should carry more weight than others, and let's face it, being cited in the Spring '96 issue of *Social Text* (pages 217–252, to be precise) isn't exactly a literary feather in your cap. But taking into account the quality of citations is only half the answer in search.

$$\vec{h}=\lambda A\vec{a}$$
$$\vec{a}=\lambda A^{t}\vec{h}$$

Because compared to the neatly governed world of scientific publishing, the Internet is completely insane. Fluid. Volatile. Heterogeneous. Awash in anonymity. Replete with conflicting agendas. So counting inbound links isn't enough. Not even close. To search effectively in these circumstances, you have to don some serious math goggles and take a look at the big picture.

THE ALGORITHM SEES GALAXIES, BUT IT'S BLIND AS A BAT.

The heavy hitters of search all use the same mathematically myopic approach—counting links back to authoritative Web pages. But the only way to tell what's really going on is to take a step back and look for patterns in the sites that point back to authorities. And when you do, you quickly see that there is another layer to the puzzle—sites that point to more than one authority, or hub pages, if you will. These hubs and their surrounding authorities form little galaxies of relevant information, something that makes the hair stand up on the back of any self-respecting searchophile's neck. It's the difference between checking out the Big Dipper from a lawn chair in your back yard and peering into Fornax with Hubble's Ultra Deep Field. But an algorithm that could detect these galaxies would be virtually impossible to pull off, since it would have to assess both inbound and outbound information, and continually calculate the relationship between the two, in real time.



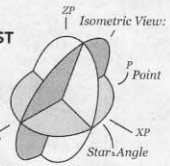
THE ALGORITHM IS RELATIVELY SIMPLE, IF YOU'RE SOME KIND OF SAVANT.

It works like this. For each search query, an index G of Web pages is found. For each page p, you associate a non-negative authority weight $a(p)$ → $a=AT\vec{h}$ and a non-negative hub weight $h(p)$. → $\vec{h}=Aa$ This will lead you to the rather obvious conclusion that when p points to lots of pages with big a values, it should get a big h value (inverse weighted popularity). And when p is pointed to by lots of pages with big h values, it should get a big a value (weighted popularity). From here, you simply fire up an iterative singular value decomposition operation and wrap things up by banging out an orthonormal basis of eigenspace for each and obtaining the eigenvectors for the matrices in question. That's it.

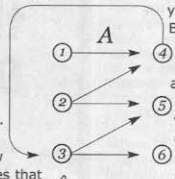
IT'S A GOOD THING ROBERT FROST NEVER WROTE AN ALGORITHM.

Taking the road less traveled is fine if you're stumbling around the New England countryside, being whimsical or whatever. But when you're searching online, that kind of thing gets you eaten by wolves. Because dismissing where others have gone can quickly get you lost in a forest of irrelevant results. But while you are learning from the Algorithm, the Algorithm is learning too. It studies the way anonymous groups of users search and forms an aggregate view of which results those users find the most valuable. This sends relevance through the roof and gets you to your desired destination without the slightest hint of lupine intercession. Sure, "The Road Traveled Every Five Minutes" would make a lousy poem, but it makes a gorgeous piece of code.

THE ALGORITHM APPROACHES ARTIFICIAL INTELLIGENCE, BUT IT HAS NOTHING AGAINST PEOPLE NAMED SARAH CONNOR.



Yes, the Algorithm is an omniscient, evolving organism devoid of all feeling, but in no way should this freak you out. In fact, it's cause for celebration. Because the Algorithm comes in peace. It's here to revolutionize search by identifying a topic, finding experts on that topic and assessing the popularity of pages among those experts, simultaneously, in the blink of an eye, whenever you want. It's here to narrow or expand your search based on *concept*—something no other search engine can do. Never again will you wade into the perpetually updated, subject-centric world of blogs without technology that actually comprehends subjects. The Algorithm knows that Usher Syndrome is transmitted by an autosomal recessive gene, not a subwoofer. And never again will you get "results" consisting merely of ten blue links, rather than the rich aggregate of images, video, conceptually related search topics and pure expert insight the Algorithm delivers.



THE ALGORITHM UNDERSTANDS THAT COLLECTIVE WISDOM IS NOT NECESSARILY COLLECTED FROM EVERYONE.

Based solely on the number of participants, the Web is undoubtedly the world's largest source of pure wisdom. But this doesn't mean there is wisdom inherent in every participant or every page. The Algorithm is acutely aware of this. It realizes that somewhere between James Surowiecki's *The Wisdom of Crowds* and Charles Mackay's *Madness of Crowds* lies the sweet spot. It sees everything but knows just what to look for. It scours the convoluted expanses of cyberspace and brings back an instantaneous convergence of wisdom collected, waiting for the day you're ready.


Newsweek
March 29
$3.95
newsweek.msnbc.com
The Next Frontiers
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

Google's Original Idea

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

$\mathcal{B}_P$ = {all pages pointing to P}

$|P|$ = number of out links from P

PageRank

Google's Original Idea

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

$\mathcal{B}_P$ = {all pages pointing to P}

$|P|$ = 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

Google's Original Idea

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

$\mathcal{B}_P$ = {all pages pointing to P}

$|P|$ = 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

Google's Original Idea

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

$\mathcal{B}_P$ = {all pages pointing to P}

$|P|$ = number of out links from P

Successive Refinement

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

Iteratively refine rankings for each page

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

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

PageRank

Google's Original Idea

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

$\mathcal{B}_P$ = {all pages pointing to P}

$|P|$ = number of out links from P

Successive Refinement

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

Iteratively refine rankings for each page

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

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

$$\vdots$$

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

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

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

— $\pi_{k+1}^T = \pi_k^T \mathbf{H}$ where $h_{ij} = \begin{cases} 1/|P_i| & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$

In Matrix Notation

After Step k

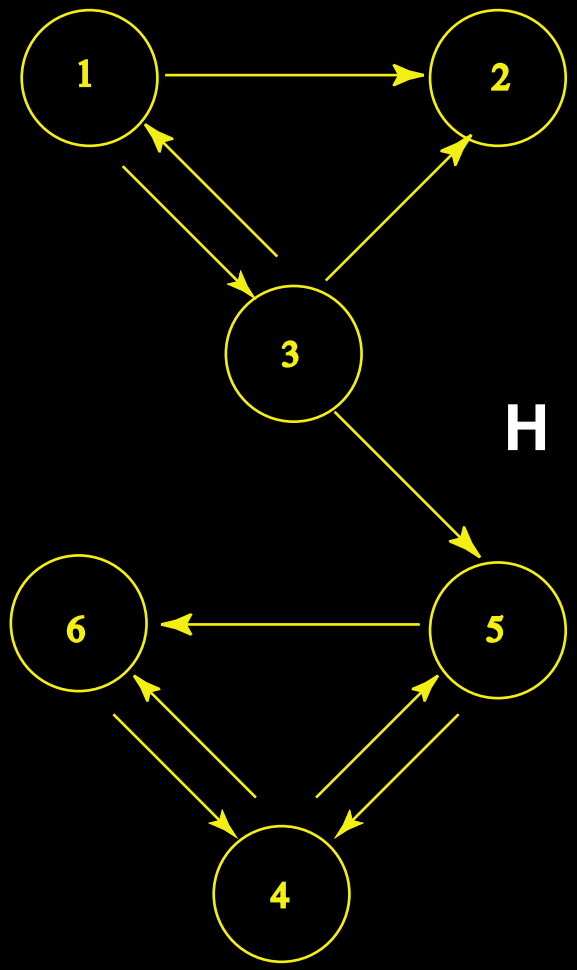
— $\boldsymbol{\pi}_k^T = [r_k(P_1), r_k(P_2), \cdots, r_k(P_n)]$

— $\boldsymbol{\pi}_{k+1}^T = \boldsymbol{\pi}_k^T \mathbf{H}$ where $h_{ij} = \begin{cases} 1/|P_i| & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$

— PageRank vector = $\boldsymbol{\pi}^T = \lim_{k \to \infty} \boldsymbol{\pi}_k^T$ = eigenvector for **H**

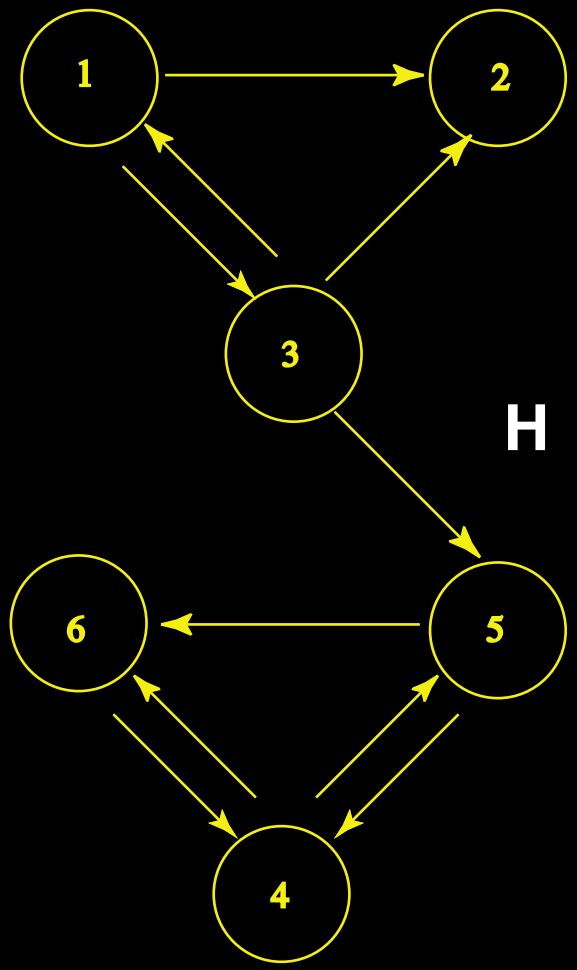
Provided that the limit exists

Tiny Web



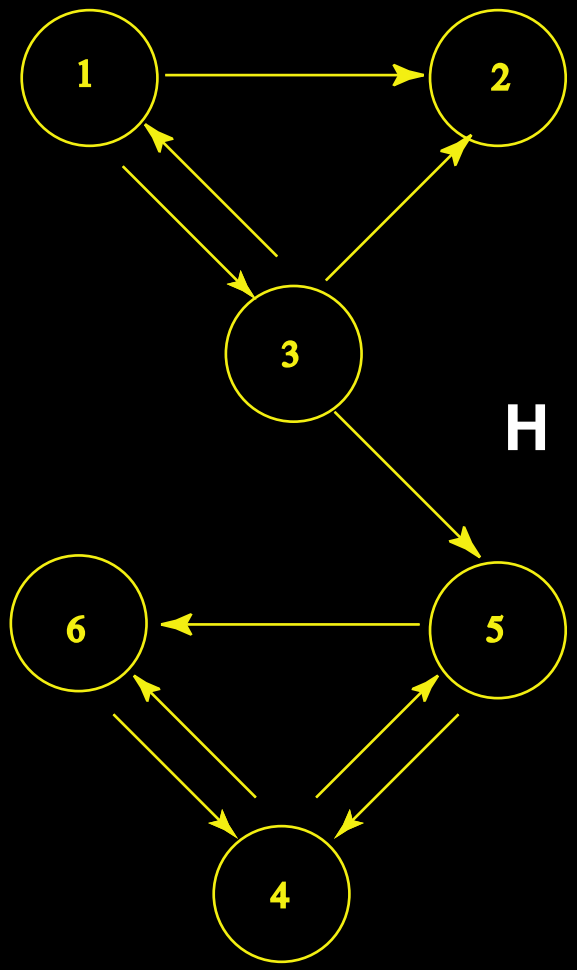
$$H = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



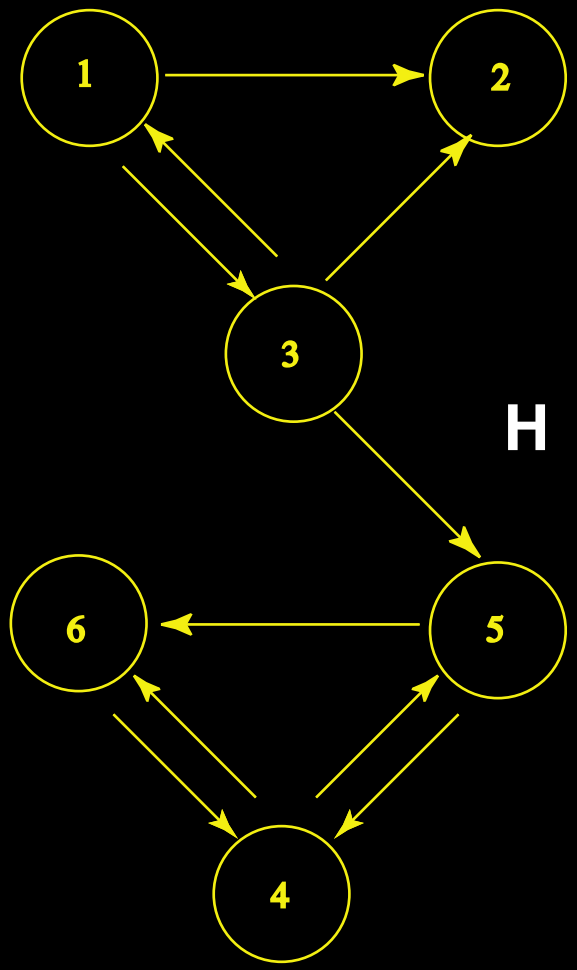
$$H = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



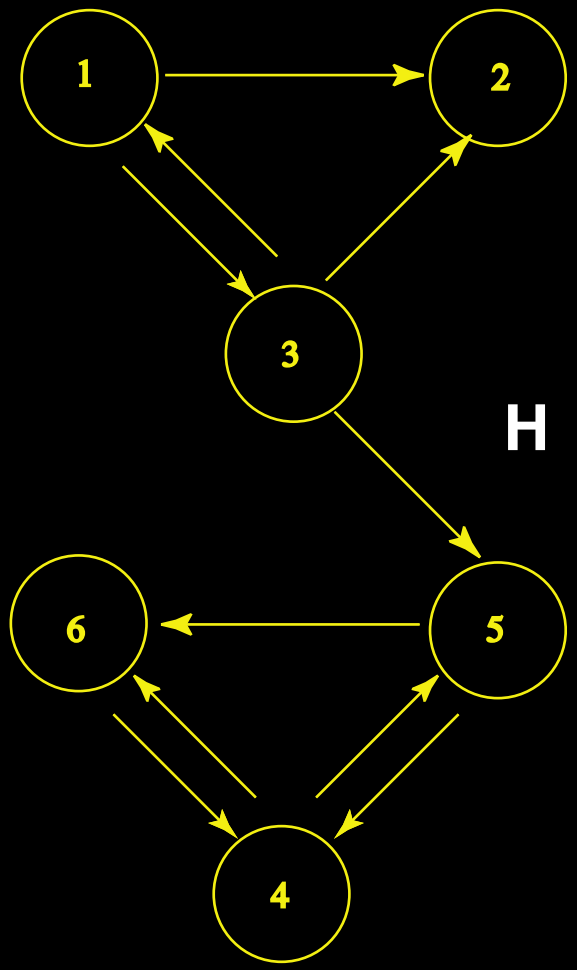
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



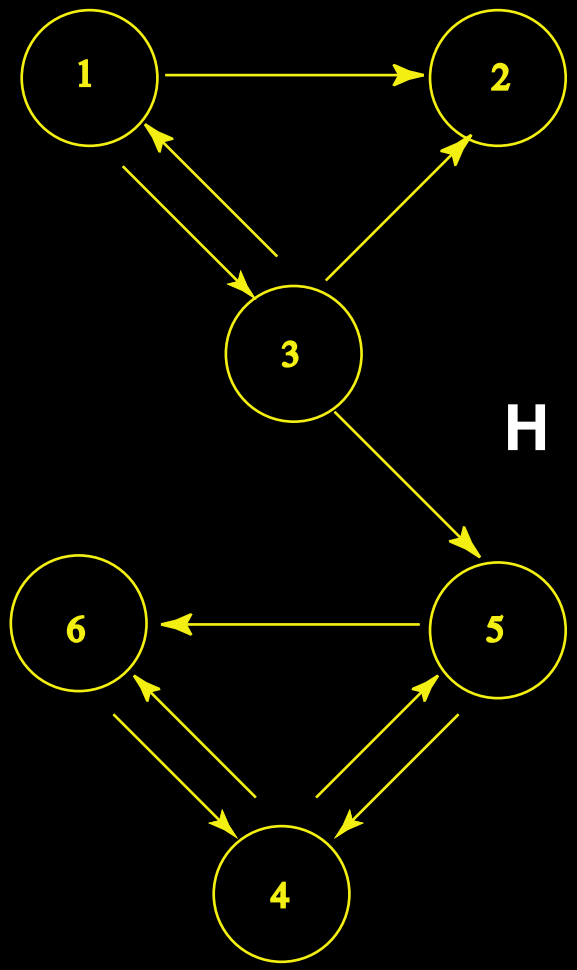
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



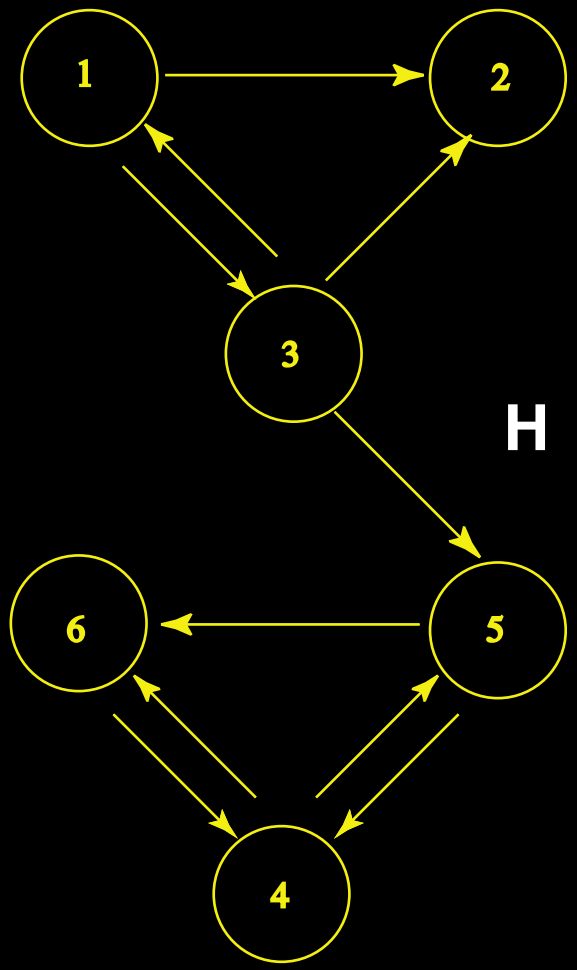
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



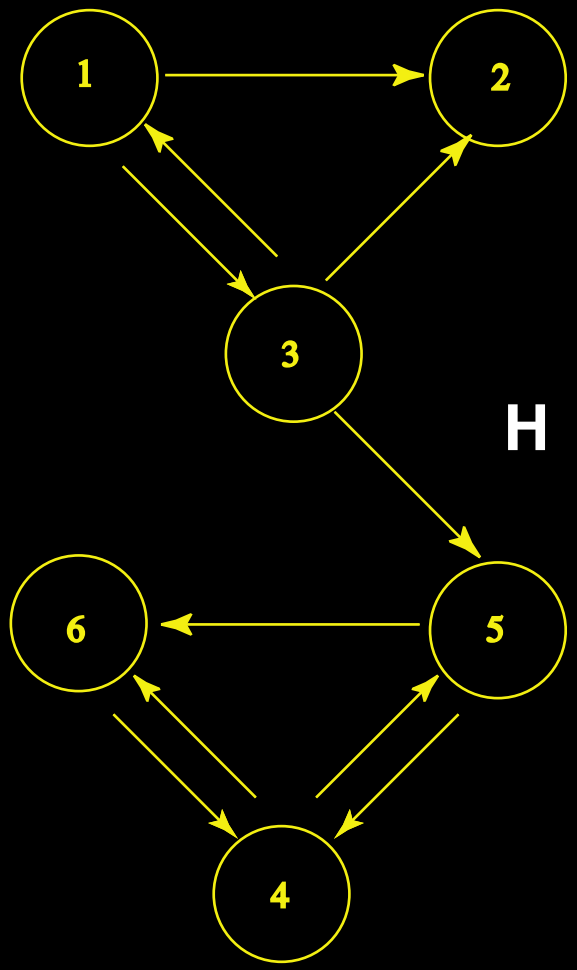
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

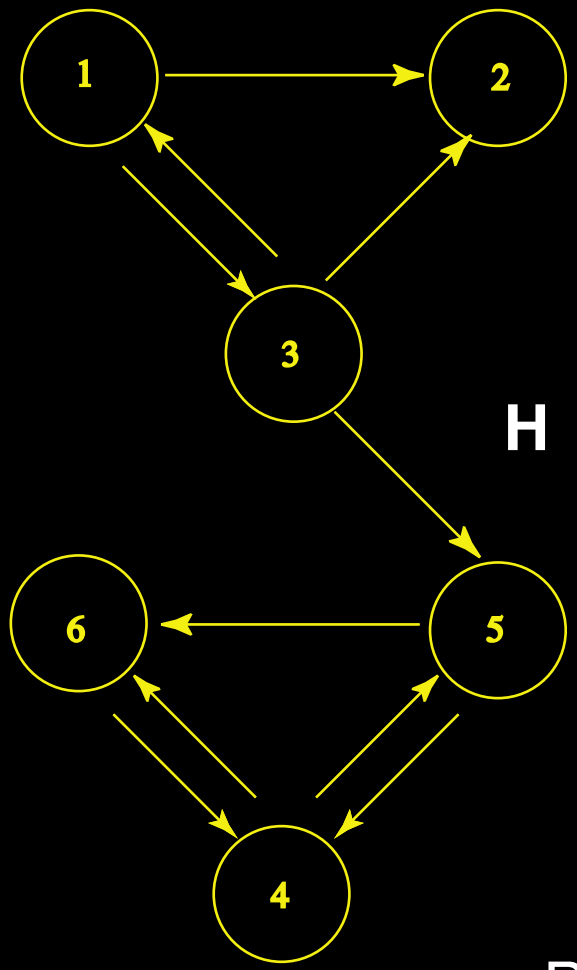
Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph

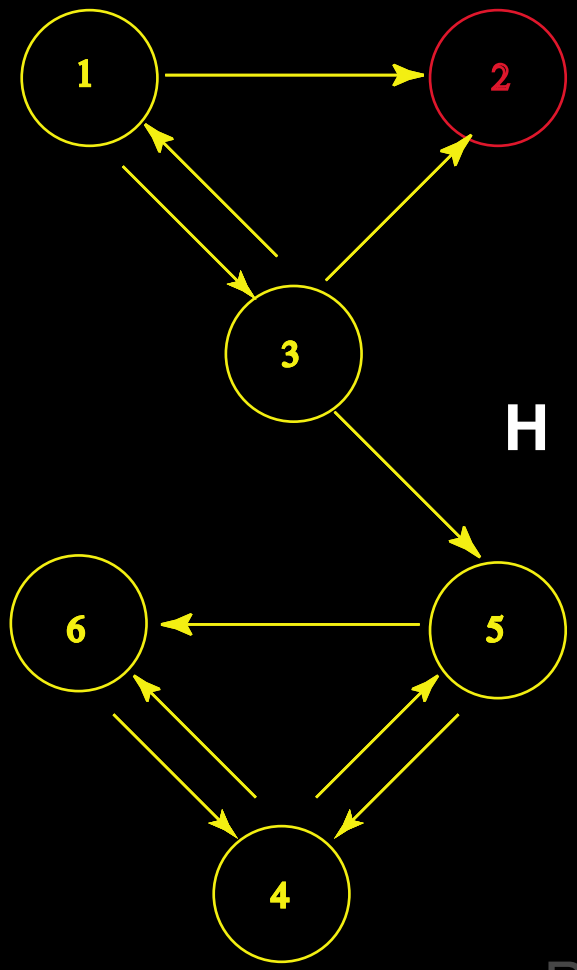
Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph
- PageRank = π_i = amount of time spent at P_i

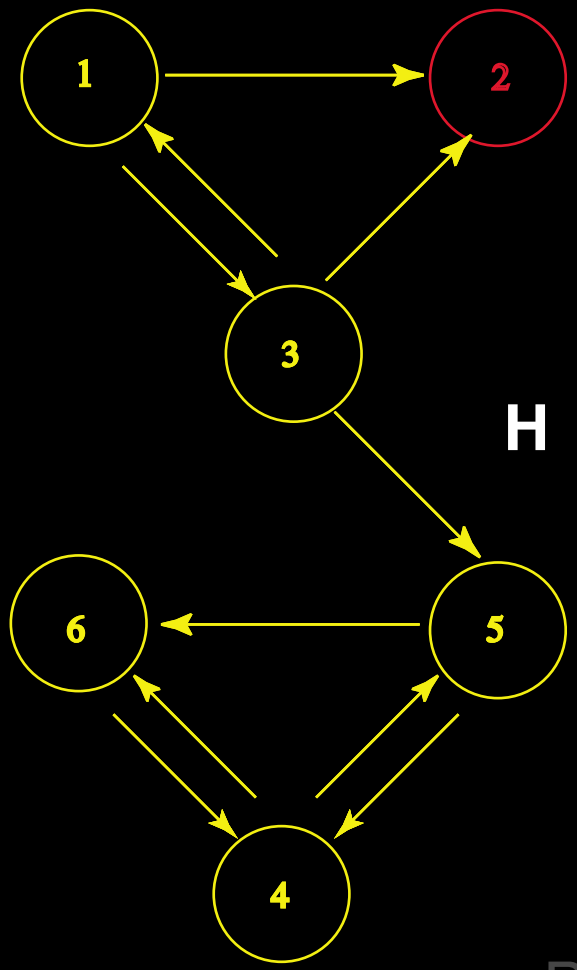
Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph
- PageRank = π_i = amount of time spent at P_i
- Dead end page (nothing to click on) — a “dangling node”

Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph
- PageRank = π_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 $\Longrightarrow$ 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 $\frac{\mathbf{e}^T}{n} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$

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

The Fix

Allow Web Surfers To Make Random Jumps

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

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

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

The Fix

Allow Web Surfers To Make Random Jumps

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

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

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

— Perron says $\exists\ \boldsymbol{\pi}^T \geq 0$ s.t. $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T\mathbf{S}$ with $\sum_i \pi_i = 1$

Nasty Problem

The Web Graph Is Not Strongly Connected

Nasty Problem

The Web Graph Is Not Strongly Connected

- i.e., S is a reducible matrix

$$\mathbf{S} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ \hline P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{ccc|ccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \left(\begin{array}{ccc|ccc} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ \hline 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{array}\right) \end{array}$$

Nasty Problem

The Web Graph Is Not Strongly Connected

— i.e., S is a reducible matrix

$$\mathbf{S} = \begin{array}{c|ccc|ccc} & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ P_2 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ \hline P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ P_6 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}$$

— Reducible $\Longrightarrow$ PageRank vector is not well defined

— Frobenius says **S** needs to be *irreducible* to ensure a unique $\pi^T > 0$ s.t. $\pi^T = \pi^T \mathbf{S}$ 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 $(P_i \to P_j \to P_i)$

- The powers $\mathbf{S}^k$ fail to converge

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \to P_j \to P_i)$

- The powers $\mathbf{S}^k$ fail to converge
- $\pi_{k+1}^T = \pi_k^T \mathbf{S}$ fails to convergence

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \to P_j \to P_i)$

- The powers $\mathbf{S}^k$ fail to converge
- $\pi_{k+1}^T = \pi_k^T \mathbf{S}$ fails to convergence

Convergence Requirement

- Perron–Frobenius requires $\mathbf{S}$ to be primitive

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \to P_j \to P_i)$

- The powers $\mathbf{S}^k$ fail to converge
- $\pi_{k+1}^T = \pi_k^T \mathbf{S}$ fails to convergence

Convergence Requirement

- Perron–Frobenius requires $\mathbf{S}$ to be primitive
- No eigenvalues other than $\lambda = 1$ on unit circle

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \to P_j \to P_i)$

- The powers $\mathbf{S}^k$ fail to converge
- $\pi_{k+1}^T = \pi_k^T \mathbf{S}$ fails to convergence

Convergence Requirement

- Perron–Frobenius requires $\mathbf{S}$ to be primitive
- No eigenvalues other than $\lambda = 1$ on unit circle
- Frobenius proved $\mathbf{S}$ is primitive $\iff \mathbf{S}^k > 0$ for some k

The Google Fix

Allow A Random Jump From Any Page

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

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector π^T = left-hand Perron vector of $\mathbf{G}$

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector π^T = left-hand Perron vector of $\mathbf{G}$

Some Happy Accidents

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

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector — π^T = left-hand Perron vector of $\mathbf{G}$

Some Happy Accidents

- $\mathbf{x}^T\mathbf{G} = \alpha\mathbf{x}^T\mathbf{H} + \beta\mathbf{v}^T$ — Sparse computations with the original link structure
- $\lambda_2(\mathbf{G}) = \alpha$ — Convergence rate controllable by Google engineers

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector — π^T = left-hand Perron vector of $\mathbf{G}$

Some Happy Accidents

- $\mathbf{x}^T\mathbf{G} = \alpha\mathbf{x}^T\mathbf{H} + \beta\mathbf{v}^T$ — Sparse computations with the original link structure
- $\lambda_2(\mathbf{G}) = \alpha$ — Convergence rate controllable by Google engineers
- $\mathbf{v}^T$ can be any positive probability vector in $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T$

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector π^T = left-hand Perron vector of $\mathbf{G}$

Some Happy Accidents

- $\mathbf{x}^T\mathbf{G} = \alpha\mathbf{x}^T\mathbf{H} + \beta\mathbf{v}^T$ Sparse computations with the original link structure
- $\lambda_2(\mathbf{G}) = \alpha$ Convergence rate controllable by Google engineers
- $\mathbf{v}^T$ can be any positive probability vector in $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T$
- The choice of $\mathbf{v}^T$ allows for personalization

THE WALL STREET JOURNAL.

WEDNESDAY, FEBRUARY 26, 2003 - VOL. CCXLI NO. 39 - ★★★ $1.00 WSJ.com

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)

* * *

■ **Ahold's problems deepened** as U.S. authorities opened inquiries into accounting at the Dutch company's U.S. Foodservice unit.

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

(Articles on Page A2)

* * *

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

(Article on Page A3)

* * *

■ **The industrials** rebounded on

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 bomb. Gen. Franks assumed command of the war-operations center in Qatar. Allied warplanes are aggressively taking out missile sites that could threaten the allied troop buildup. (Column 4 and Pages A4 and A6)

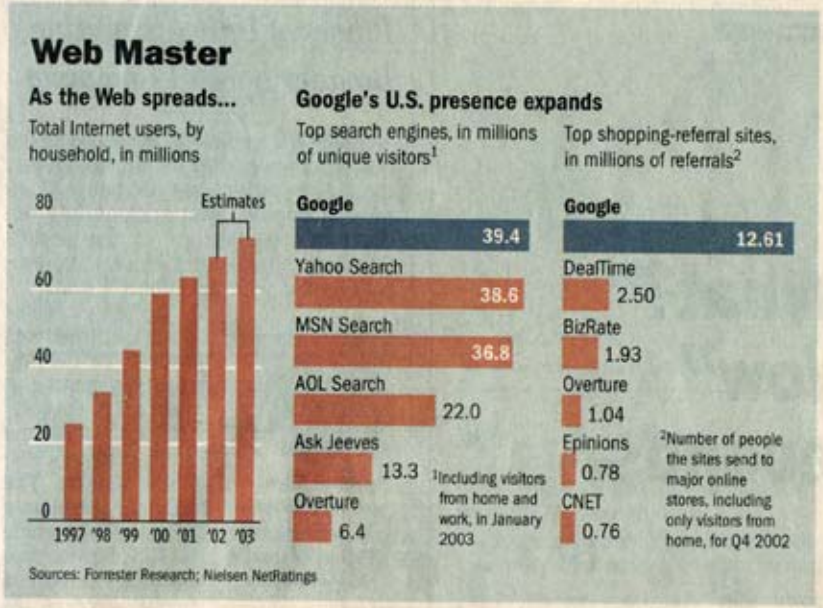
Turkey's parliament debated legislation to let the U.S. deploy 62,000 to open a northern front. Kurdish soldiers lined roads in a show of force as U.S. officials traveled into Iraq's north for an opposition conference.

* * *

■ **Powell said** North Korea hasn't restarted a reactor and plutonium-processing facility at Yongbyon, hinting such forbearance might constitute an overture. But saber rattling continued a day after a missile test timed for the inauguration in Seoul. Pyongyang accused U.S. spy planes of violating its airspace and told its army to prepare for U.S. attack. (Page A14)

* * *

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



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

By Greg Jaffe
And John D. McKinnon

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

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

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

Cat and Mouse

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

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

Exoticleatherwear Gets Cut Off

By Michael Totty
And Mylene Mangalindan

Joy Holman sells provocative leather clothing on the Web. She wants what nearly everyone doing business online wants: more exposure on Google.

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

The techniques

Web Sites Fight for Prime Real Estate on Google

Continued From First Page

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

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

Coding Trickery

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

Mr. Brin, 29, one of Google's two founders and now its president of technology, boasted to a San Francisco search-engine conference in 2000 that Google wasn't worried about having its results clogged with irrelevant results because its search methods couldn't be manipulated.

That didn't stop search optimizers from finding other ways to outfox the system. Attempts to manipulate Google's results even became a sport, called Google bombing. Pranksters would try to

creating Web sites that were nothing more than collections of links to the clients' site, called "link farms." Since Google ranks a site largely by how many links or "votes" it gets, the link farms could boost a site's popularity.

In a similar technique, called a link exchange, a group of unrelated sites would agree to all link to each other, thereby fooling Google into thinking the sites have a multitude of votes. Many sites also found they could buy links to themselves to boost their rankings.

Ms. Holman, the leatherwear retailer, discovered the consequences of trying to fool Google. The 42-year-old hospital laboratory technician, who learned computer skills by troubleshooting her hospital's

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

equipment, operates her online apparel store as a side business that she hopes can someday replace her day job.

When she launched her Exotic Leather Wear store from her home in Mesa, Ariz., she quickly learned 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 "haltertops" or "leather miniskirts." Since Google doesn't give much weight to metatags in determining its rankings, the efforts had little effect on her search results.

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

In theory, when Google encounters the AutomatedLinks code, it treats it as a legitimate referral to the other sites and counts them in toting up the sites' popularity.

Shortly after Ms. Holman signed up with AutomatedLinks in July, she read on an online discussion group that Google objected to such link arrangements. She says she immediately stripped the code from her Web pages. For a while her site gradually worked its way up in Google search results, and business steadily improved because links to her site still remained on the sites of other AutomatedLinks customers. Then, sometime in November, her site was suddenly no longer appearing among the top results. Her orders plunged as much as 80%.

Ms. Holman, who e-mailed Google and AutomatedLinks, says she has been unable to get answers. But in the last few months, other AutomatedLinks customers say they have seen their sites apparently penalized by Google. Graham McLeay, who runs a small chauffeur service north of London, saw revenue cut in half during the two months he believes his site was penalized by Google.

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

AutomatedLinks didn't respond to requests for comment. Google declined to comment on the case. But Mr. Cutts, the Google engineer, warns that the rules are clear and that it's better to follow them rather than try to get a problem fixed after a site has been penalized. "We want to return the most relevant pages we can," Mr. Cutts says. "The best way for a site owner to do that is follow our guidelines."

Crackdown

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

To stay one step ahead of the Web

homa City-based SearchKing, an online directory for hundreds of small, specialty Web sites. SearchKing also sells advertising links designed both to deliver traffic to an advertiser and boost its rankings in Google and other search results.

Bob Massa, SearchKing's chief executive, last August launched the PR Ad Network as a way to capitalize on Google's page-ranking system, known as PageRank. PageRank rates Web sites 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.

Shortly after publicizing the ad network, Mr. Massa discovered that his site suddenly dropped in Google's rankings. What's more, sites that participated in the separate SearchKing directory also had their Google rankings lowered. He filed a lawsuit in Oklahoma City federal court, claiming Google was punishing him for trying to profit from the company's page-ranking system.

A Google spokesman won't comment on the case. In its court filings, Google said it demoted pages on the SearchKing site because of SearchKing's attempts to manipulate search results. The company has asked for the suit to be dismissed, arguing that the PageRank represents its opinion of the value of a Web site and as such is protected by the First Amendment.

"The big search engines determine the laws of how commerce runs," says Mr. Massa, who is persisting with the lawsuit even though the sites have had their page rankings partly restored. "Someone needs to demand accountability."

Google is taking steps that many say could satisfy businesses trying to boost their rankings. Google has long sold sponsored links that show up on the top of many search-results pages, separate from the main listings. Last year, the company expanded its paid-listings program, so that there are now more slots where sites can pay for a prominent place in the results. Many sites now are turning to advertising instead of tactics to optimize their rankings.

Home Depo Amid First

By Chad Terhun

ATLANTA—**Home Depot I** fiscal fourth-quarter earnin 3.4% on disappointing sales.

Speaking to investors a analysts, the company's ch chief executive, Bob Na Home Depot is prepared t dissatisfied customers and competitive challenge from val with remodeled stores, i ventory and improved custo

The nation's largest ho ment retailer said net income ter ended Feb. 2 decreased to or 30 cents a share, from $71 30 cents a share, a year earli 2% to $13.21 billion from $13.4 first quarterly sales decline i ny's 24-year history. Home the latest quarter was a week a year earlier. Using compar periods, the company said qu increased 5% and net income

Same-store sales, or sal open at least a year, decline quarter. Home Depot said st last month offset a disastrou and helped the retailer avoi estimate that same-store sale as much as 10%. In 4 p.m Stock Exchange composite tr Depot shares rose 66 cents t

Fiat Patria Is Set to Bec

By Alessandra Ga

ROME—Umberto Agnelli named **Fiat** SpA chairman on ping into the driver's seat as th glomerate works on an 11th-h ing of its unprofitable car un

Mr. Agnelli, the 68-year-o Fiat patriarch Gianni Agne last month, was widely exp over from current chair Fresco, later this year. But who has served as chairma

Personalization is Coming

The Wall Street Journal April 25, 2007

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. 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

Please turn to page D8

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 Mathematics
- **That Is Changing The World**