Google's PageRank

and Beyond

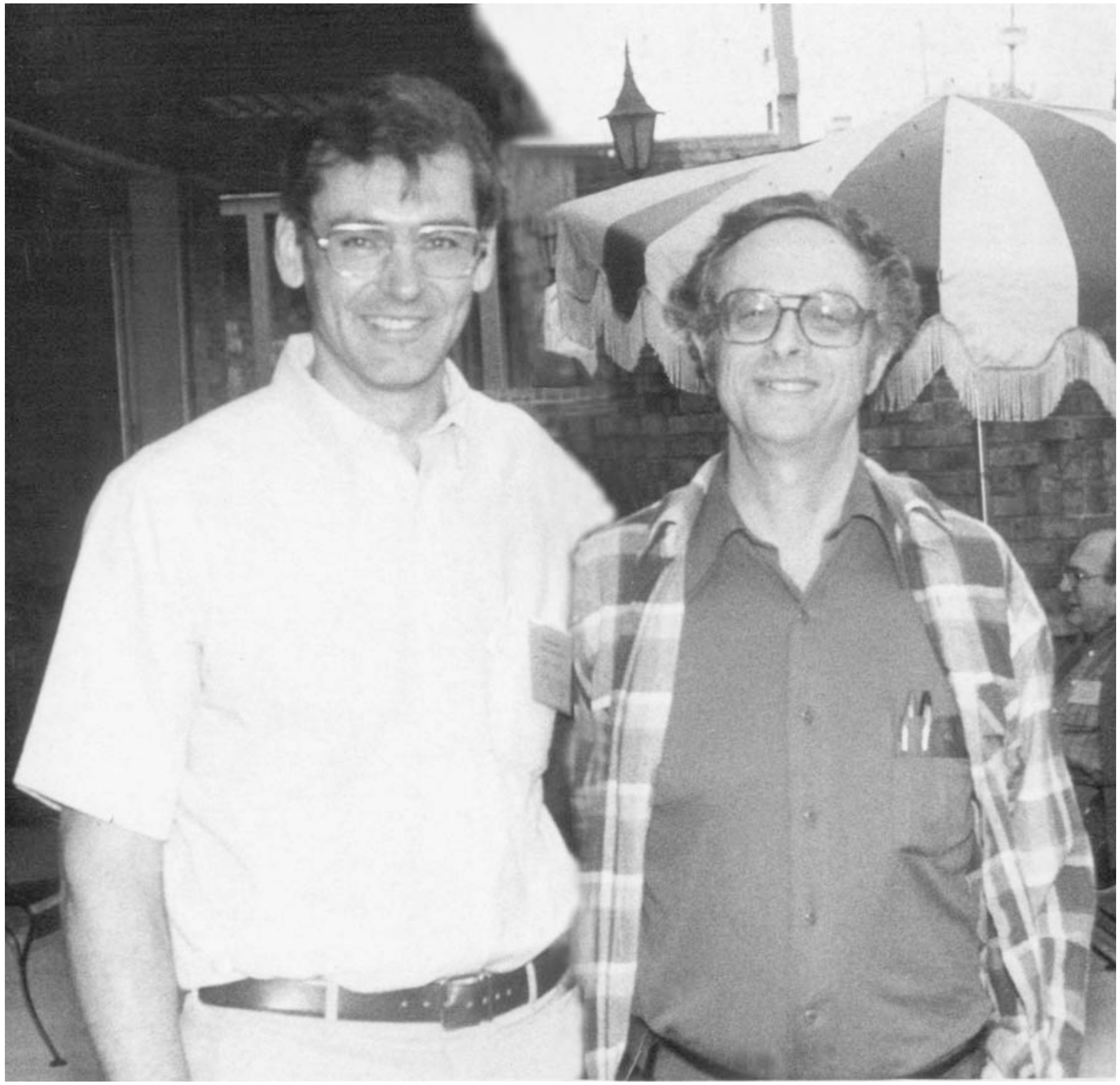
Carl Meyer

Department of Mathematics
North Carolina State University
Raleigh, NC

University of Wisconsin

May 4, 2007

“Why I Love Perron–Frobenius”
1982



Beautiful mathematics eventually tends to be useful,
and useful mathematics eventually tends to be beautiful.

Short History of IR

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

B. C.	cave paintings
12th cent. A.D.	invention of paper, monks in scriptoriums
1450	Gutenberg's printing press
1700s	Franklin's public libraries
1872	Dewey's decimal system
	Card catalog
1940s-1950s	Computer

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

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

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?

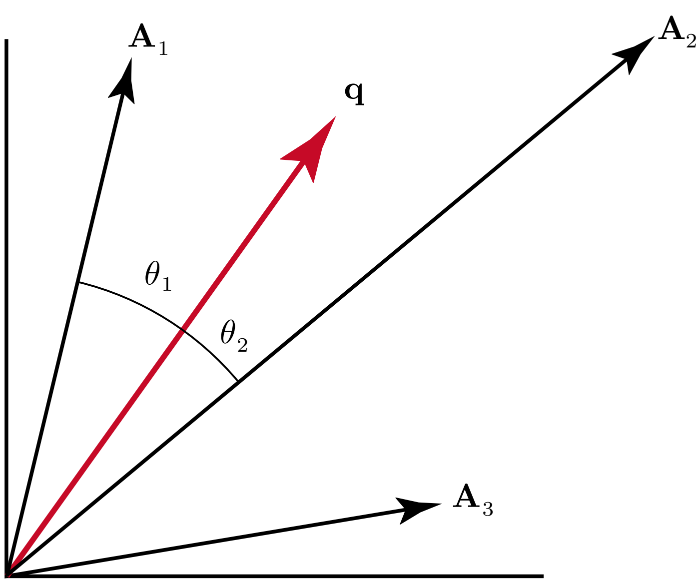
Query Matching

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 $\mathbf{A}_i$?



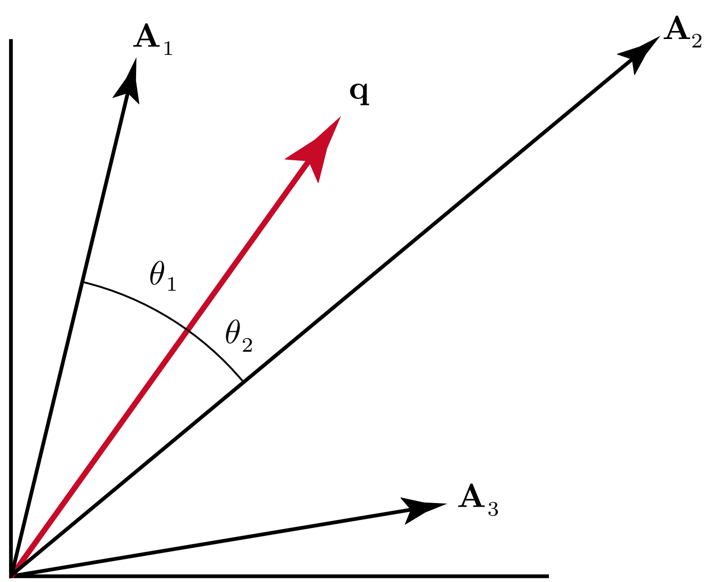
Query Matching

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 $\mathbf{q}$ to each column $\mathbf{A}_i$?



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

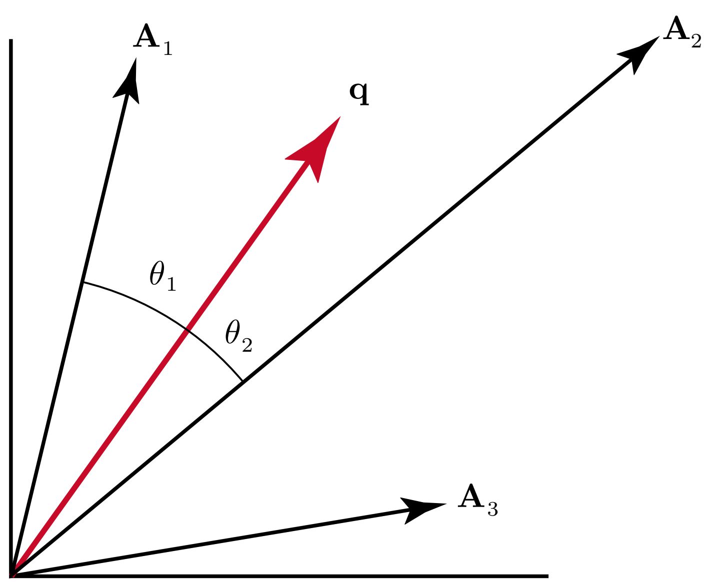
Query Matching

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 $\mathbf{q}$ 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 Fourier expansion of A

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

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

Latent Semantic Indexing

Use a 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 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 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 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

"Best" mathematical solution

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

Strengths & Weaknesses

Pros

- Finds hidden connections

Strengths & Weaknesses

Pros

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

Strengths & Weaknesses

Pros

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

Strengths & Weaknesses

Pros

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

Cons

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

Strengths & Weaknesses

Pros

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

Cons

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

Strengths & Weaknesses

Pros

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

Cons

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

Strengths & Weaknesses

Pros

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

Cons

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

Web Stats

Different from other document collections

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

Web Stats

Different from other document collections

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

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!

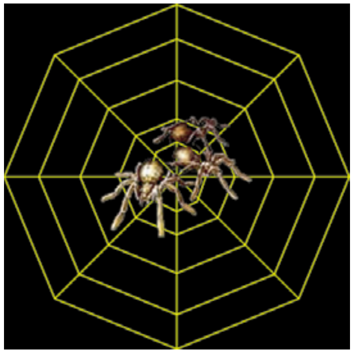
Web Stats

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

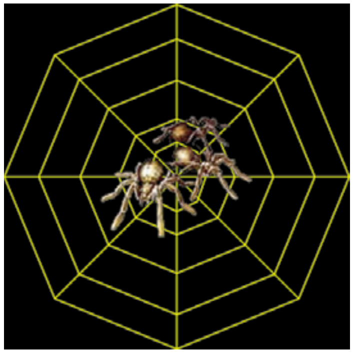
Web Crawlers



Software robots gather web pages

Web Search Components

Web Crawlers



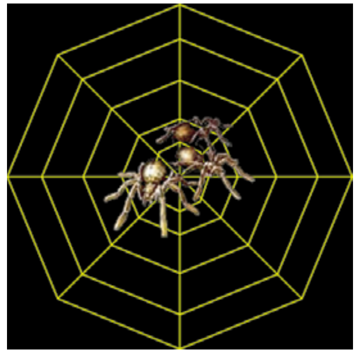
Software robots gather web pages

Doc Server

Stores docs and snippits

Web Search Components

Web Crawlers

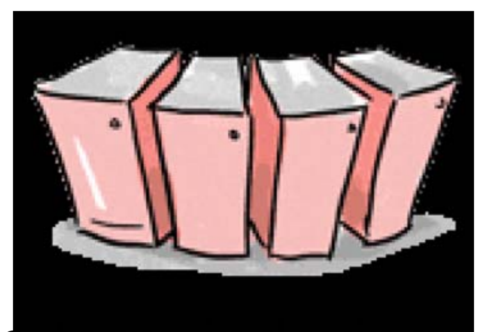


Software robots
gather web pages

Doc Server

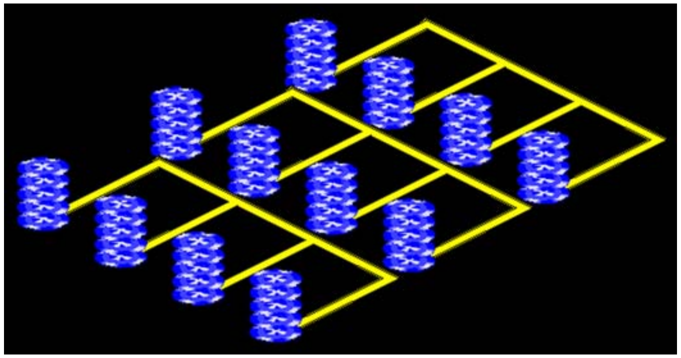
Stores docs
and snippits

Index Server



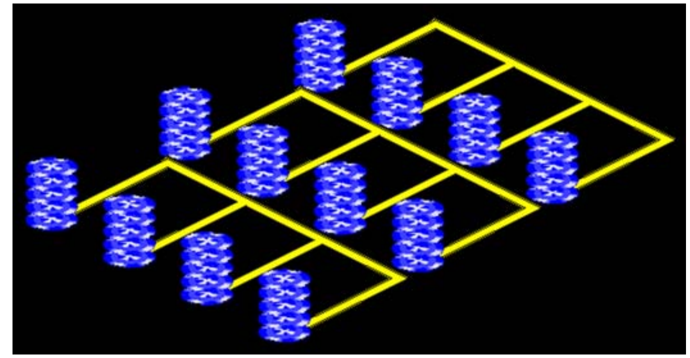
Scans pages and does term indexing
Terms $\longrightarrow$ Pages (similar to book index)

The Ranking Module



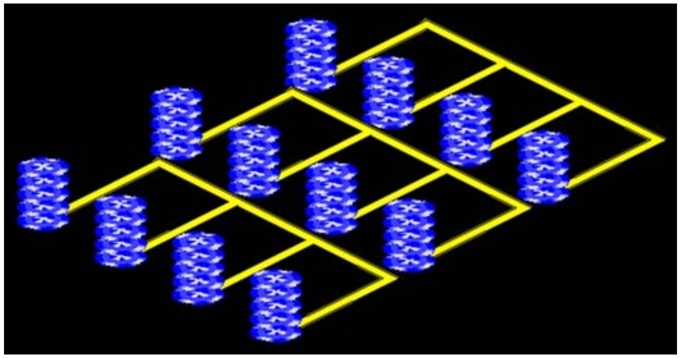
- Measure the importance of each page

The Ranking Module



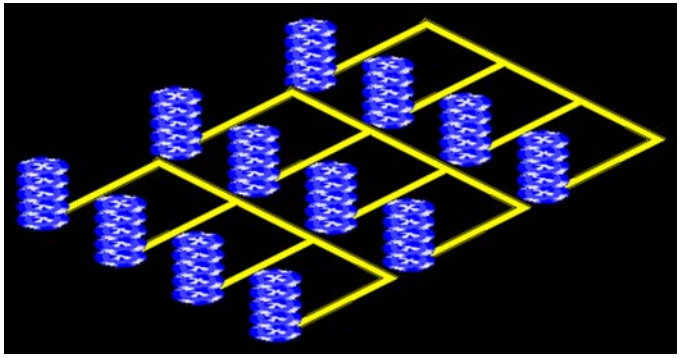
- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations

The Ranking Module



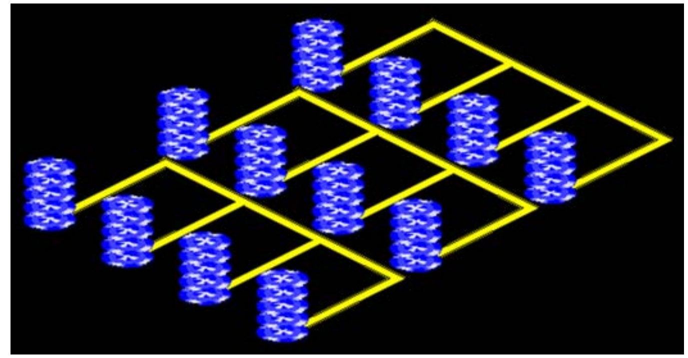
- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
- Compute these measures off-line long before any queries are processed

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

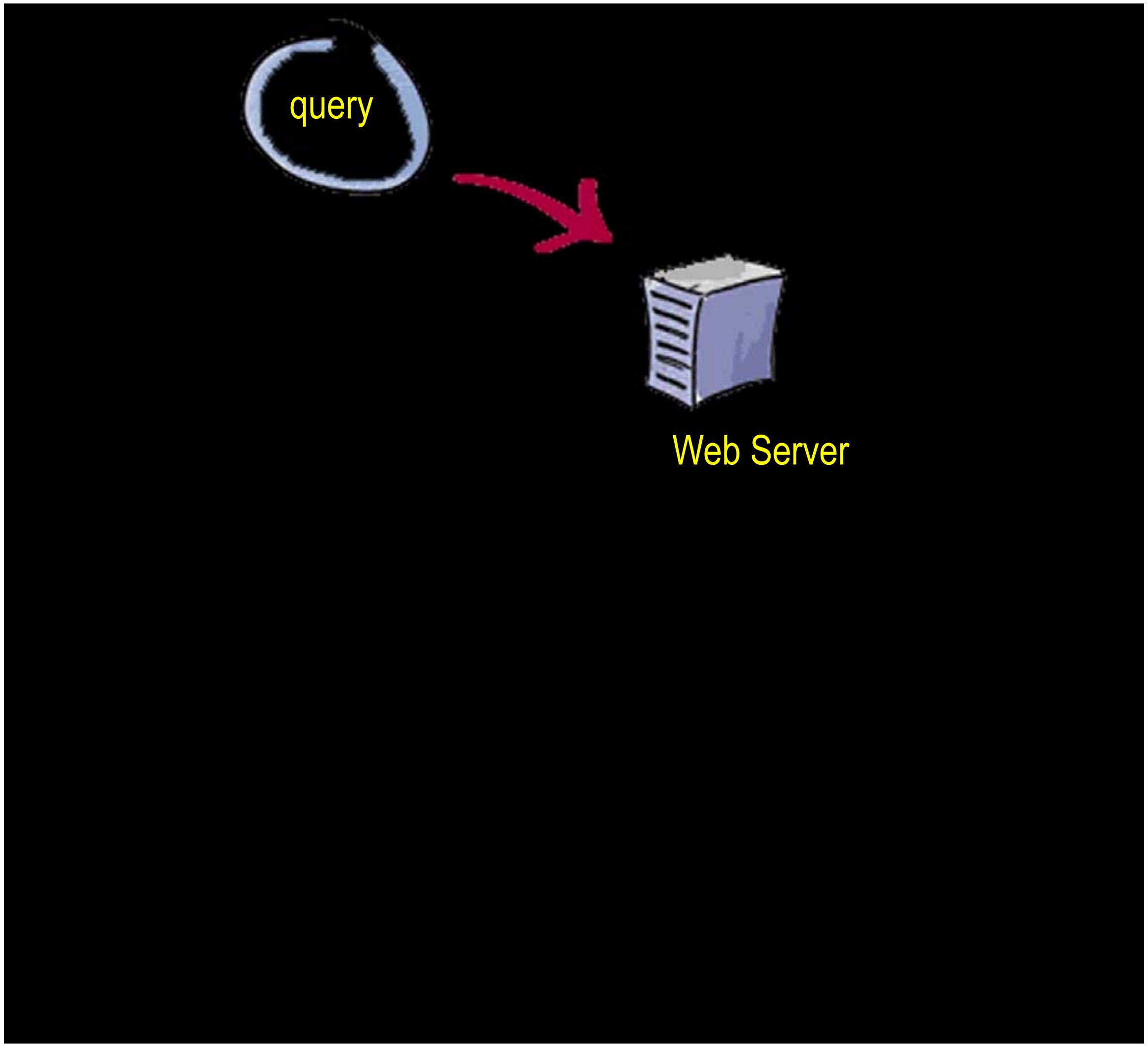
The Ranking Module



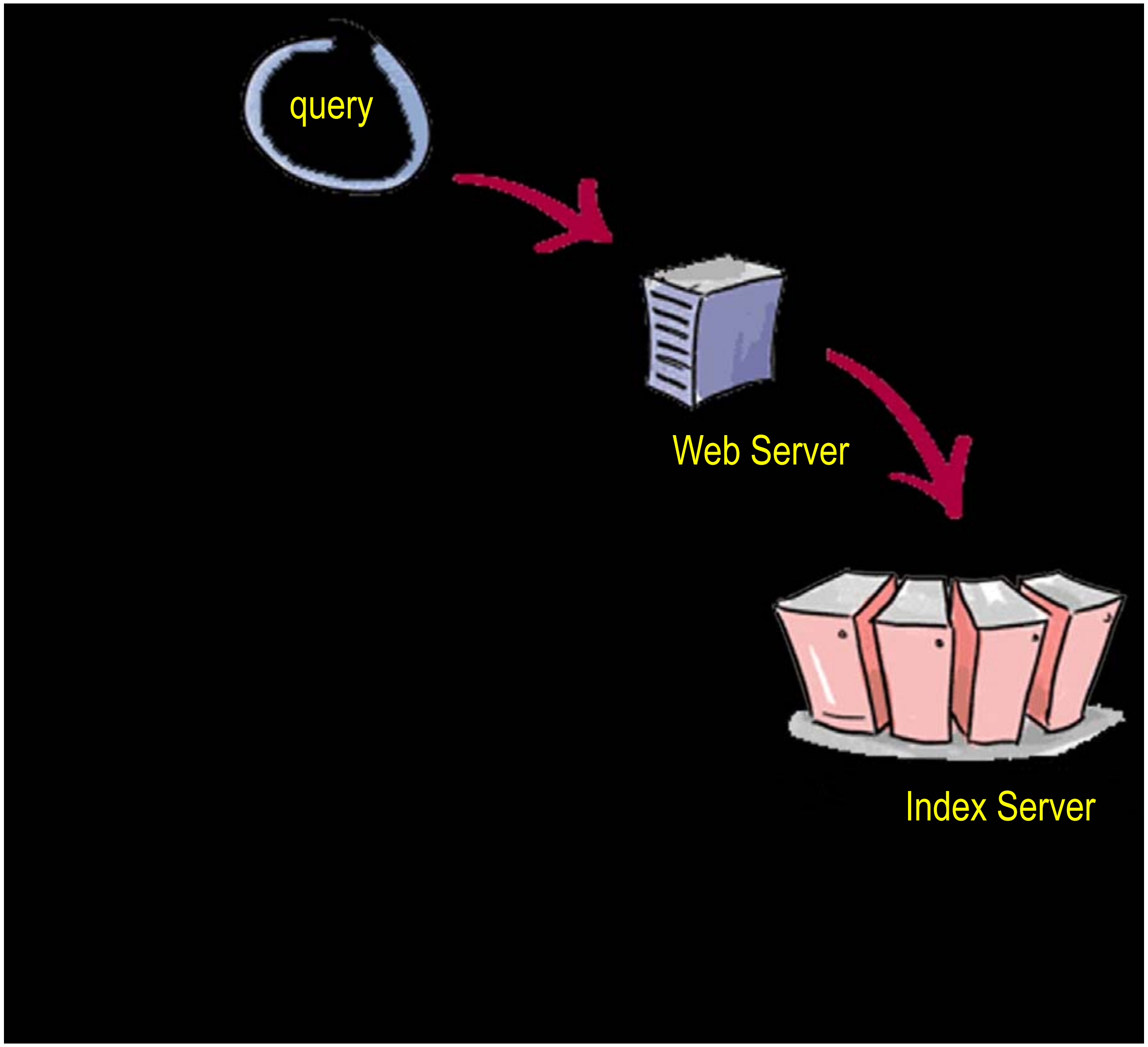
- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
- Compute these measures off-line long before any queries are processed
- Google's PageRank© technology distinguishes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

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

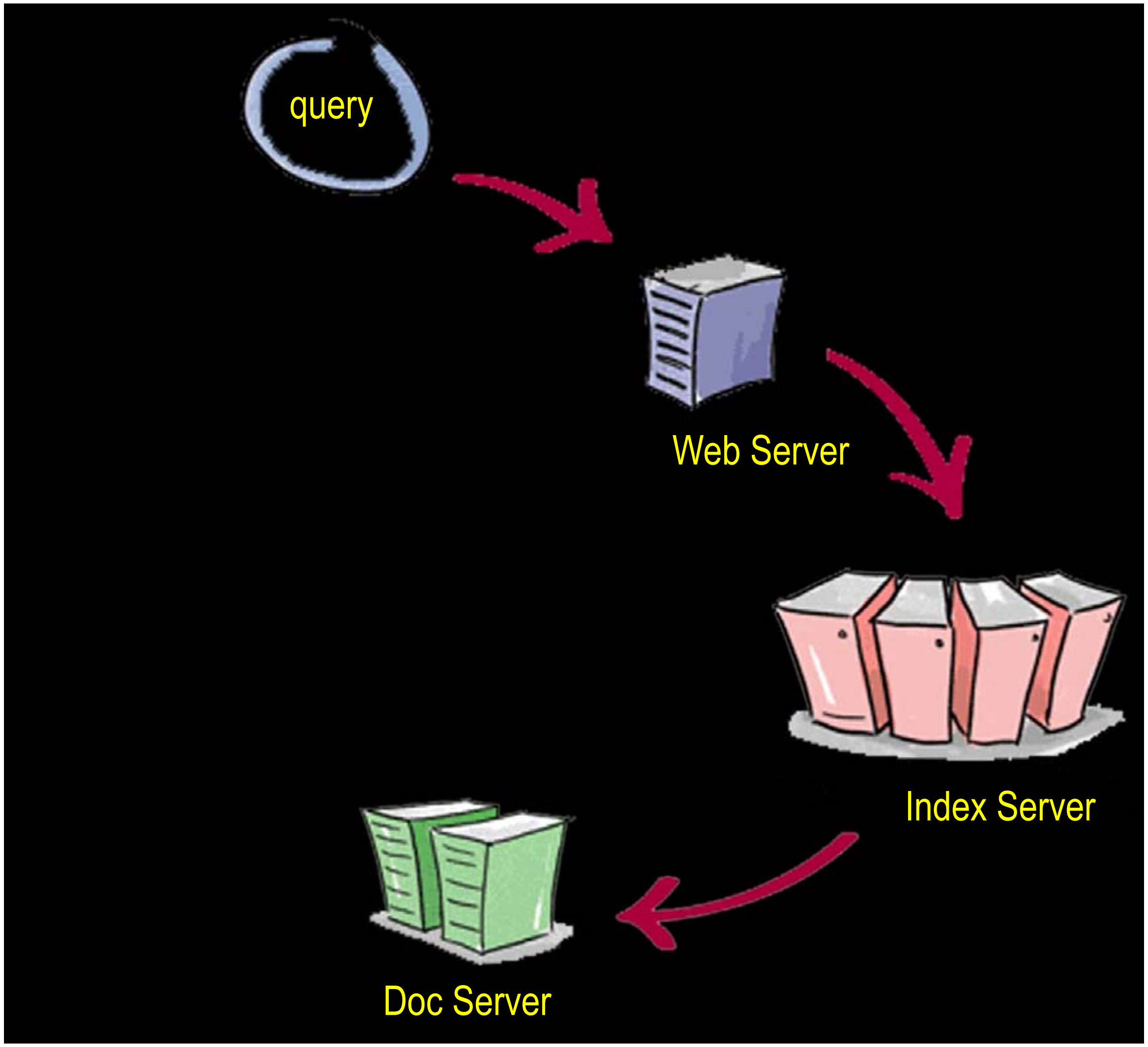
The Process



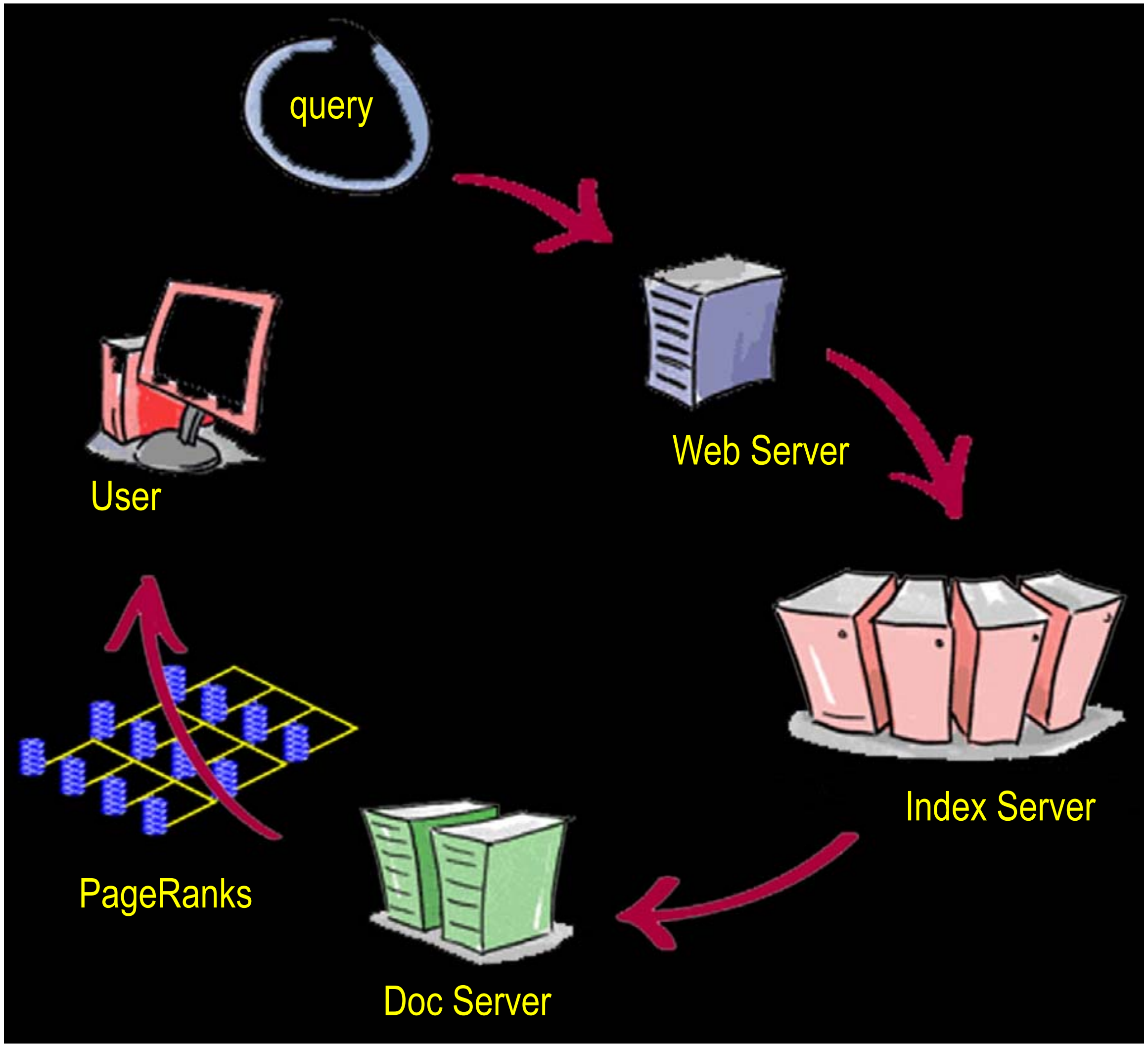
The Process

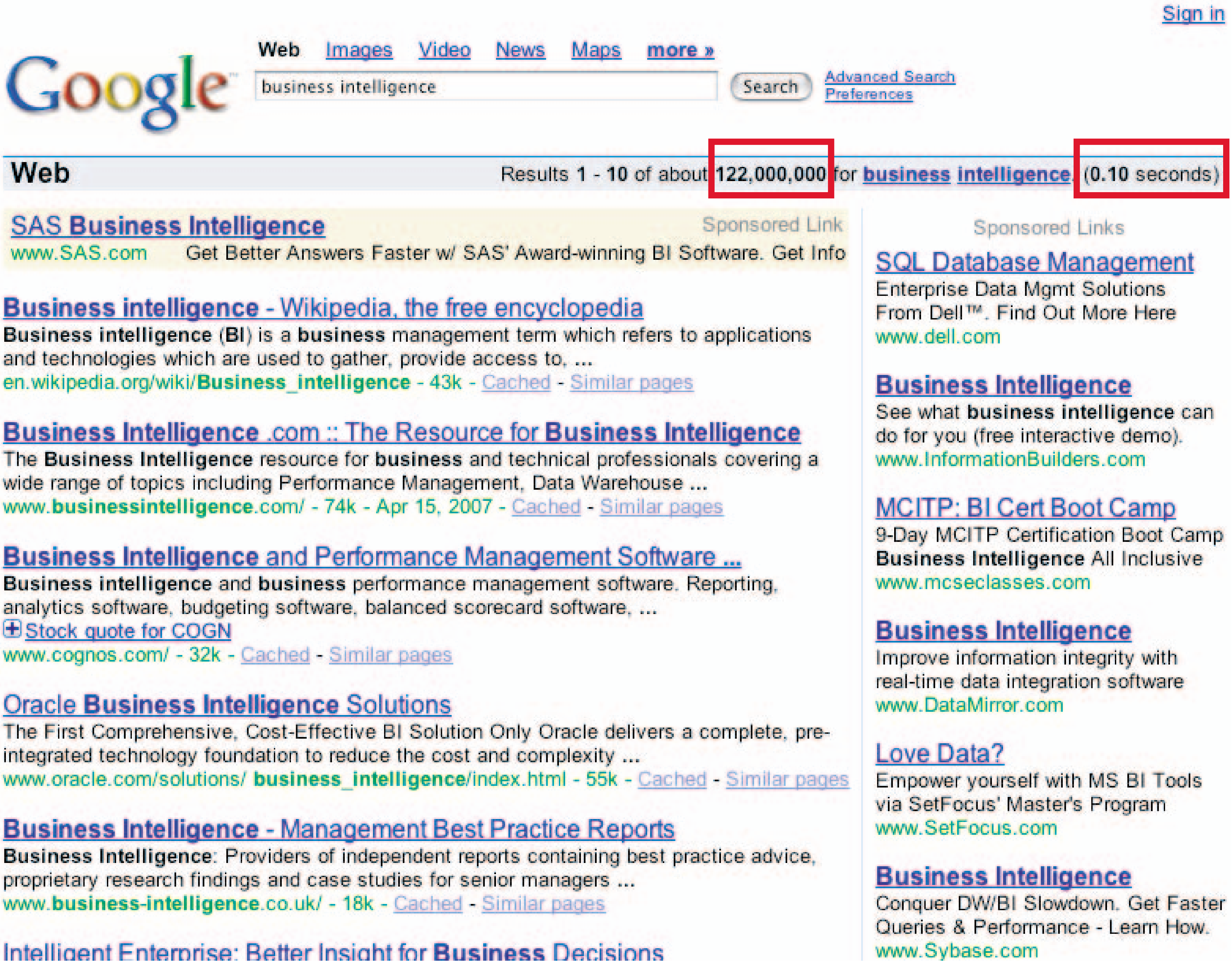


The Process



The Process

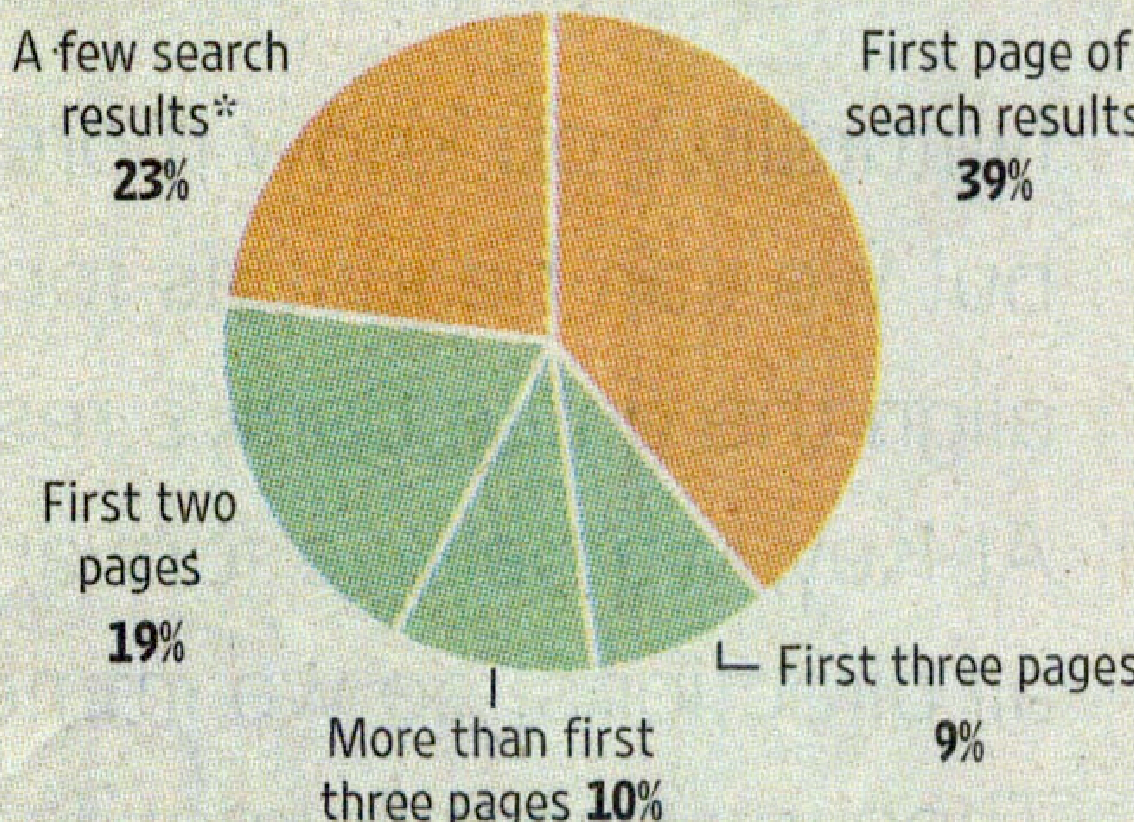



Sign in
Web Images Video News Maps more »
Google
business intelligence
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)
SAS Business Intelligence
Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software.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

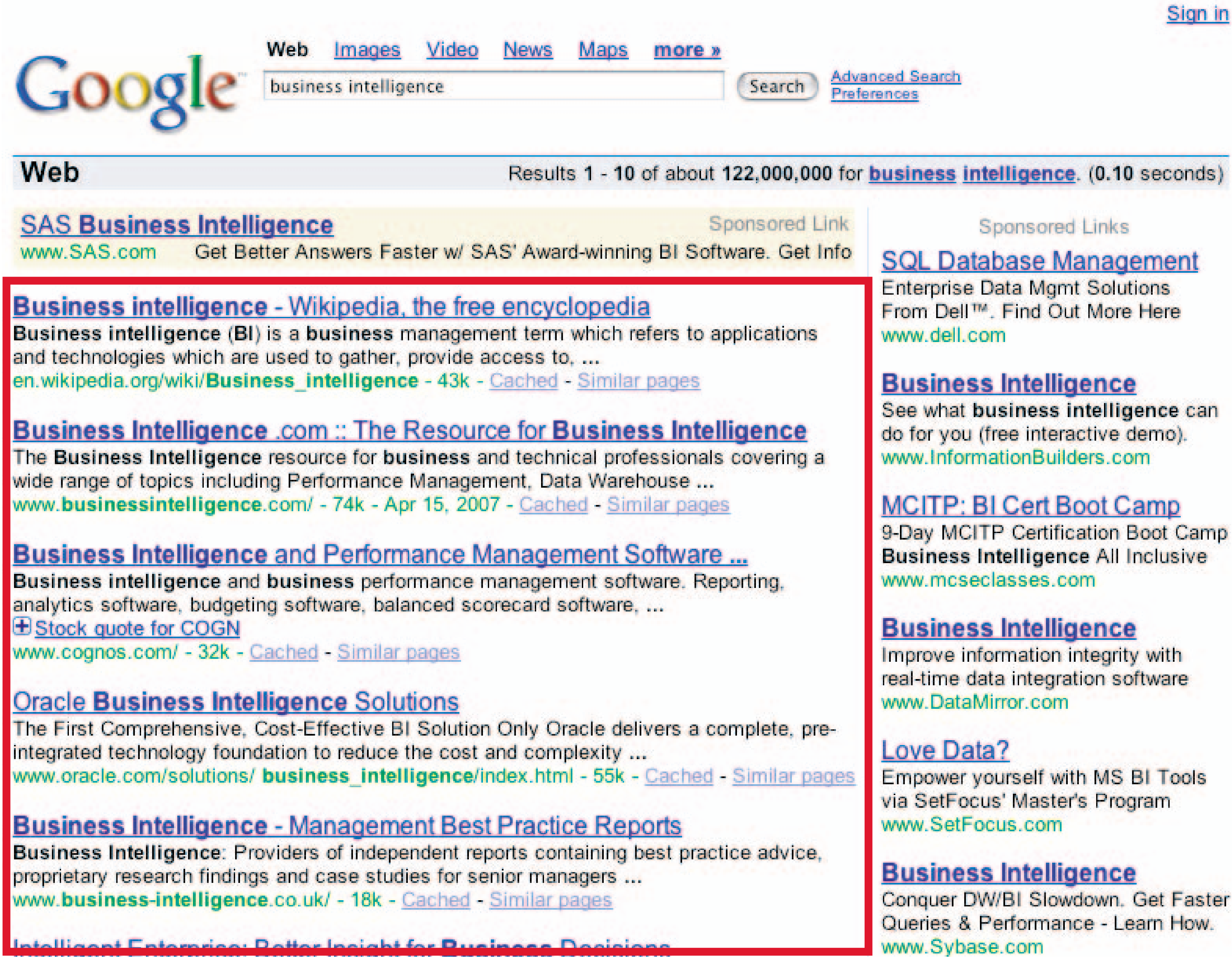
Amount of Internet search results that Web surfers typically scan before selecting one

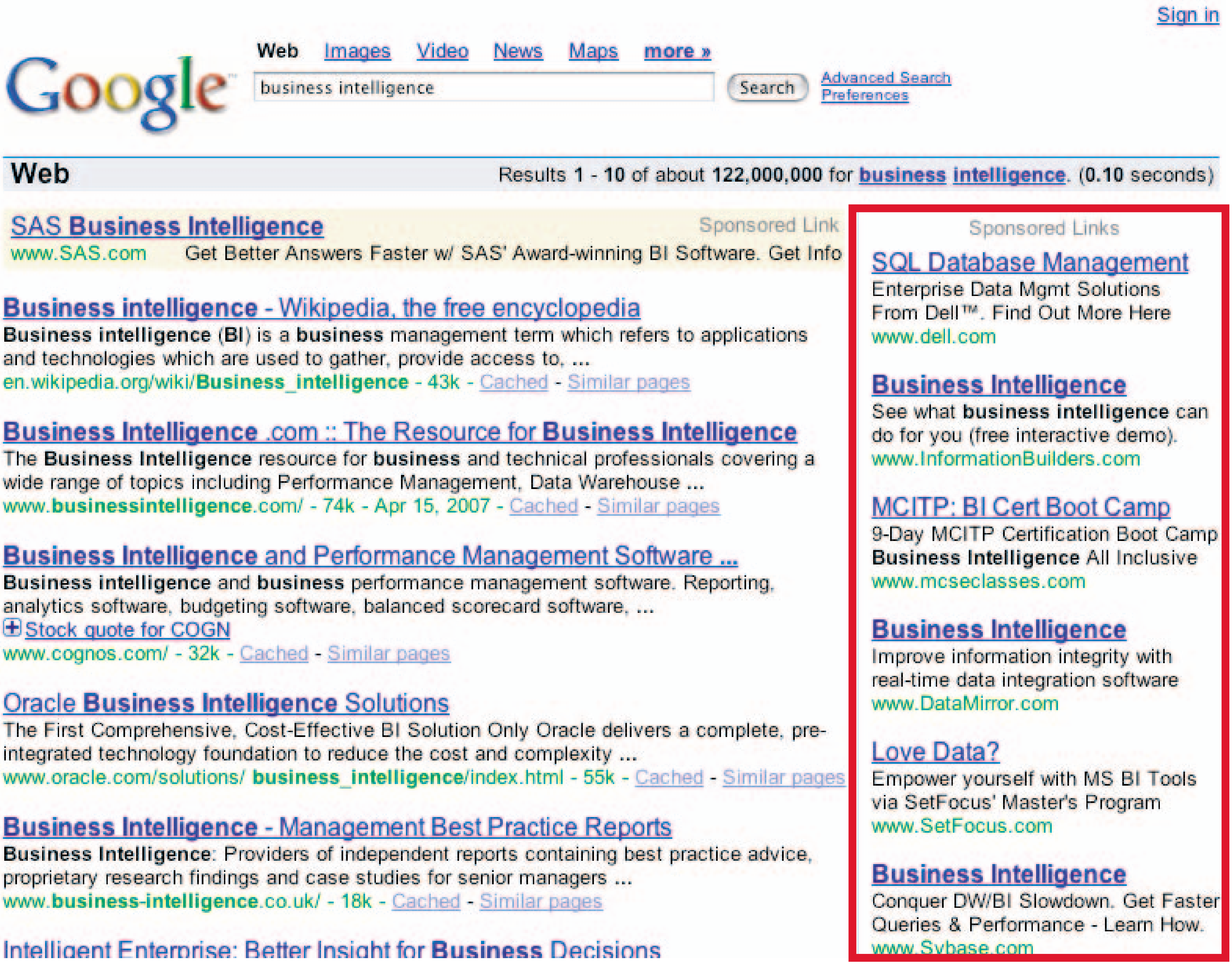


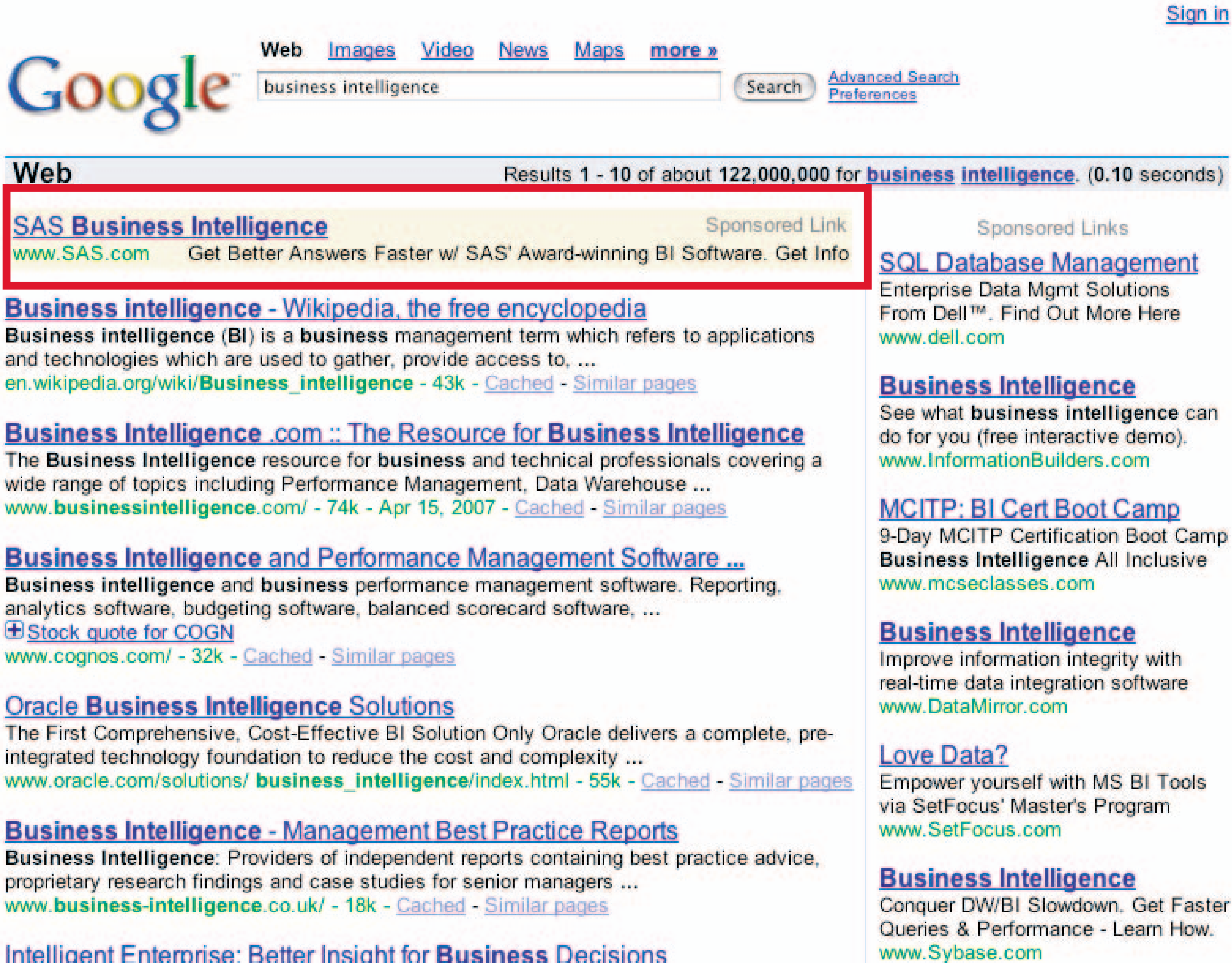
*Top results without reading through the whole page

Note: Sample size is 2,369 people

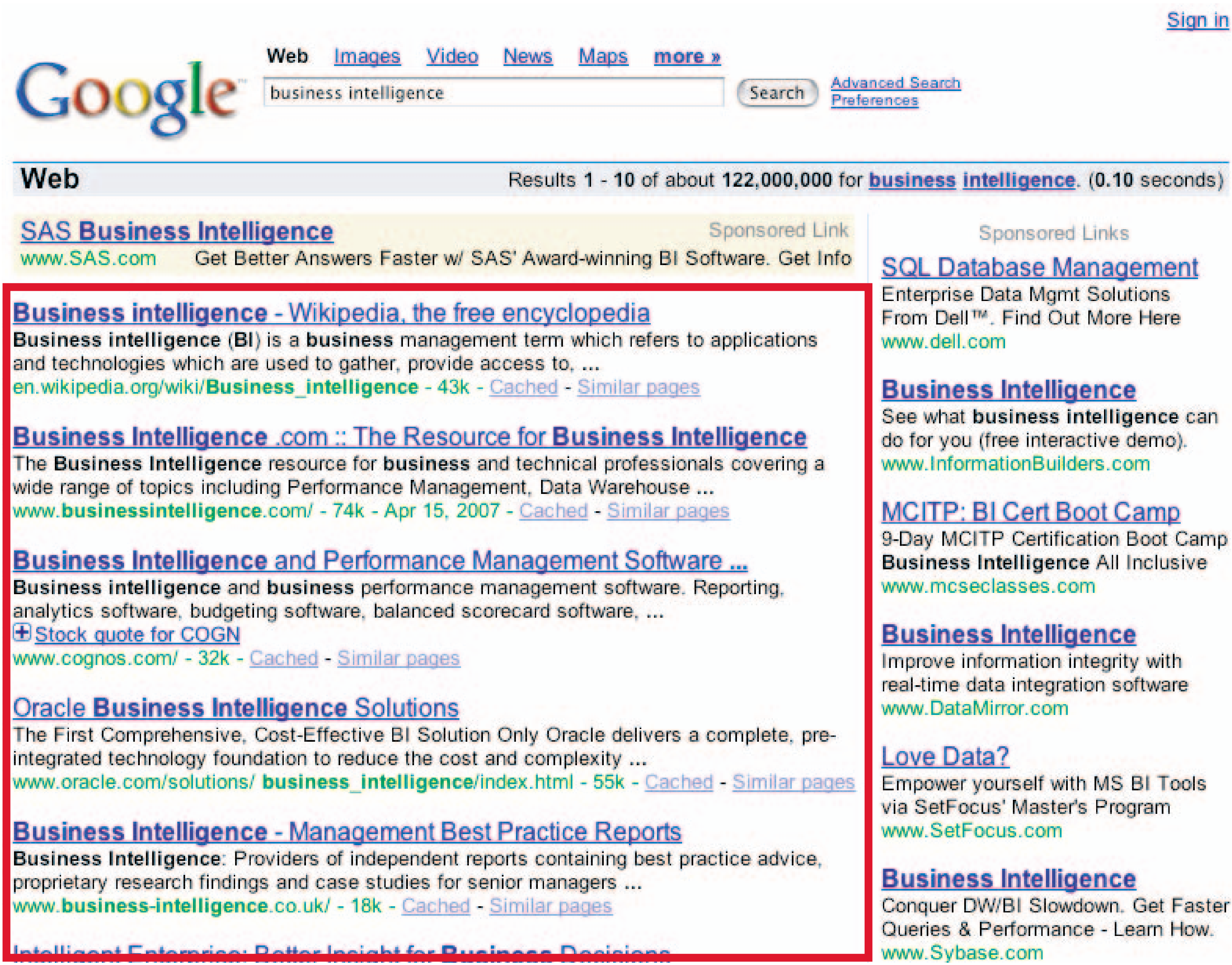
Sources: JupiterResearch; iProspect


Sign in
Web Images Video News Maps more »
Google
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
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
Web Images Video News Maps more »
Google
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



Sign in
Web Images Video News Maps more »
Google
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


Sign in
Web Images Video News Maps more »
Google
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
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 18, 2007

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.

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

In millions of dollars. Analysts' consensus estimate ... Thomson Financial; may exclude certain items or discont... operations.

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

ads such / amid cor / panies. / ama will / will see / the comi / company / tionship / "We / progress / Officer / with ana / revenue / gun impr / ter. She / moderat / generate / ing for e / company / overall / ing start / to the se

April 20, 2007

Google profit up 69 percent

Rival Yahoo's earnings fall

THE ASSOCIATED PRESS

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

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

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

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

In the latest demonstration of its clout, Google earned $1 billion, or $3.18 cents per share, during the first three months of the year. That compared with net income of $592.3 million, or $1.95 per share, in the same period last year. It was also the second consecutive quarter in which Google earned $1 billion — nearly as much money as the nation's largest newspaper publisher, Gannett, made all of last year.

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

Quarterly revenue reached a new company high of $3.66 billion, a 63 percent increase. After subtracting advertising commissions and other payments to its partners, Google's revenue totaled $2.53 billion.

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

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

As usual, Google's financial power flowed from its search engine. That ubiquitous tool has become the hub of the Web's largest marketing network and appears to be getting even better at identifying the right ads to show with its search results, helping elicit more revenue-generating clicks.

The paid clicks on the ads within Google's vast network increased 52 percent for the first quarter, compared with year-ago levels. And more of the clicks are occurring on Google's own Web sites, increasing the company's profits because the revenue doesn't have to be shared with an advertising partner.

Although Google has been trying to develop revenue beyond the Internet, online advertising continues to produce virtually all of its profit. The company is expected to become even more dominant in that business with last year's $1.76 billion acquisition of online video leader YouTube and its recently announced $3.1 billion deal to buy Internet ad distributor DoubleClick.

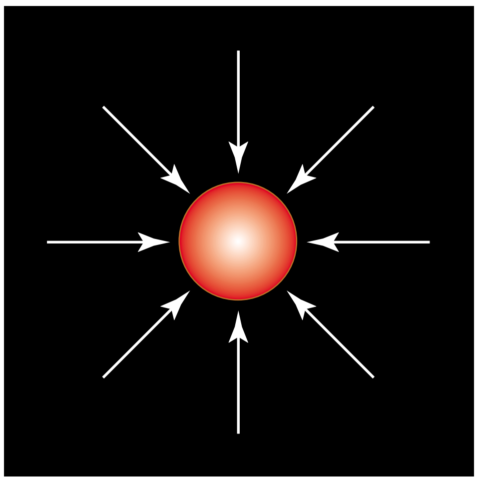
Although Google still isn't making money from YouTube, the site is "going gangbusters," co-founder Larry Page said during Thursday's conference call.

Besides buying other companies, Google is investing heavily to accommodate its growth by hiring workers and adding computer capacity at its data centers. The company spent $597 million on capital expenditures in the first quarter and hired 1,564 employees to expand to 12,238 people.

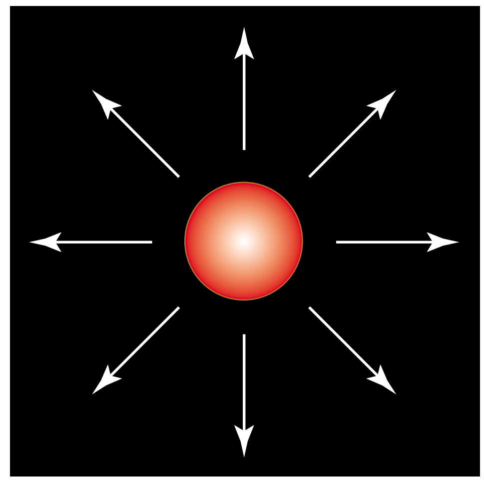
Even so, the company ended the quarter with $11.9 billion in cash.

How To Measure "Importance"

Landmark Result Paper

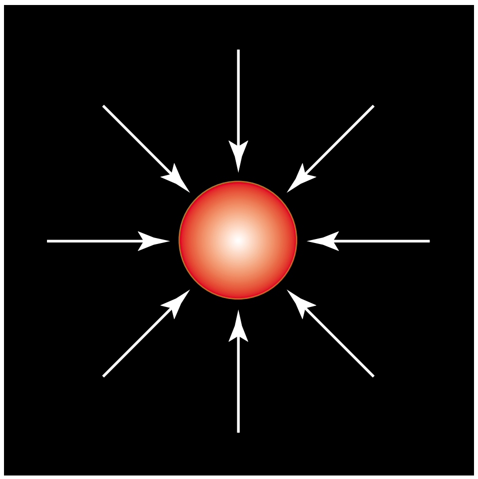


Survey Paper—Big Bib



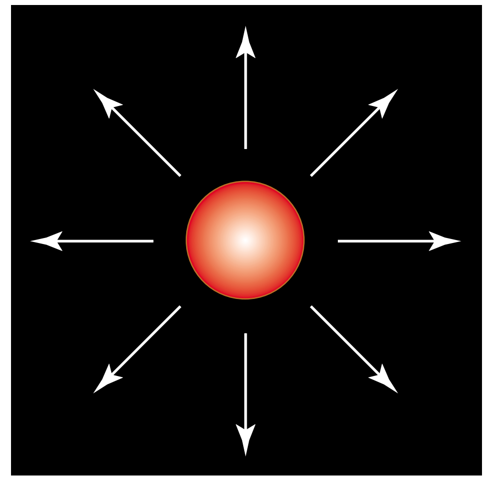
How To Measure “Importance”

Landmark Result Paper



Authorities

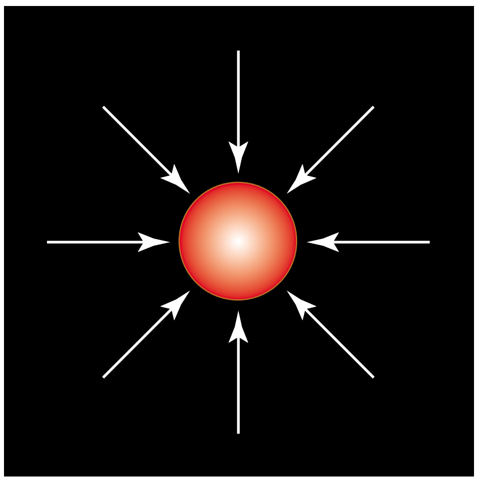
Survey Paper—Big Bib



Hubs

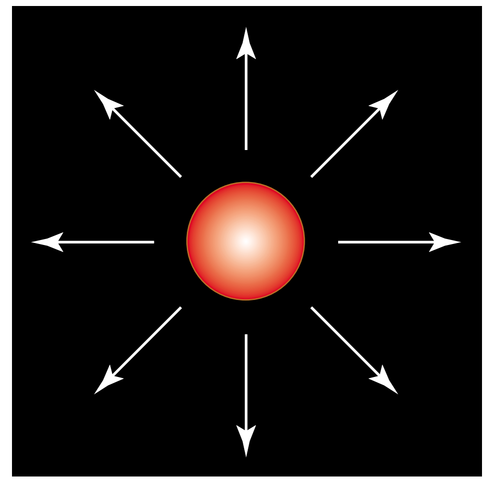
How To Measure “Importance”

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 $\Rightarrow$ $\mathbf{h}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

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 $\Rightarrow$ $\mathbf{h}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
- 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 $P_i \Rightarrow \mathbf{h}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
- 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

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

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

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

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

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

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

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

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

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

$\ddots$

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)

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

$\ddots$

Combined Iterations

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

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

$\ddots$

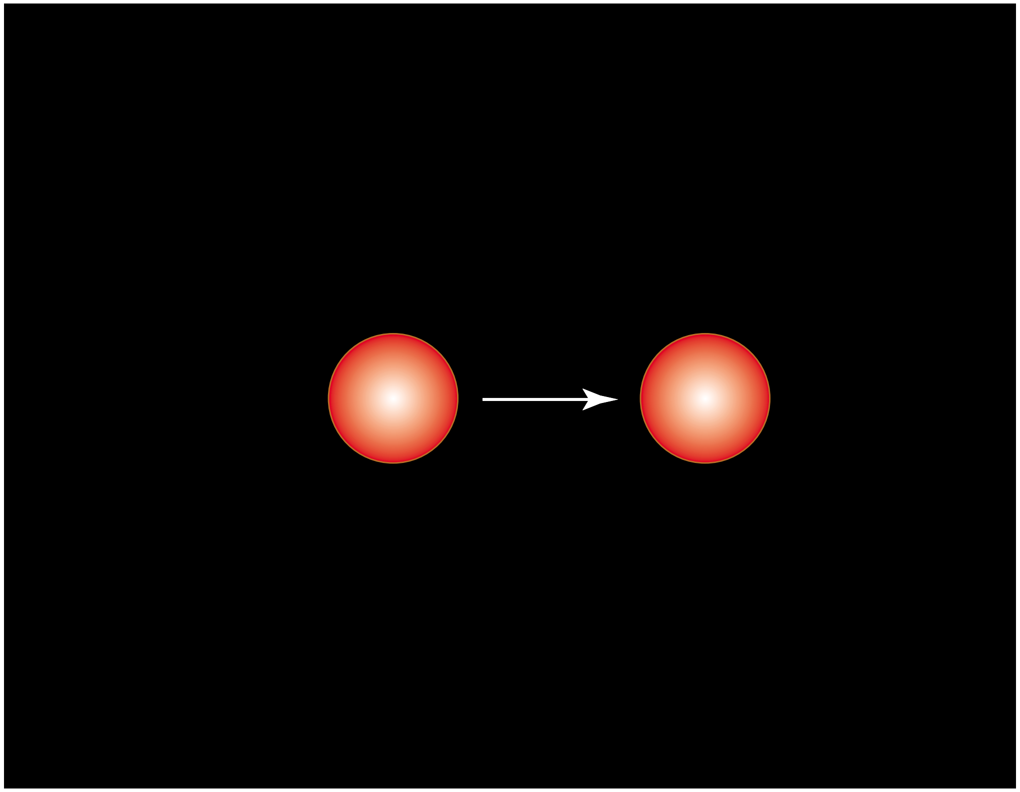
Combined Iterations

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

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

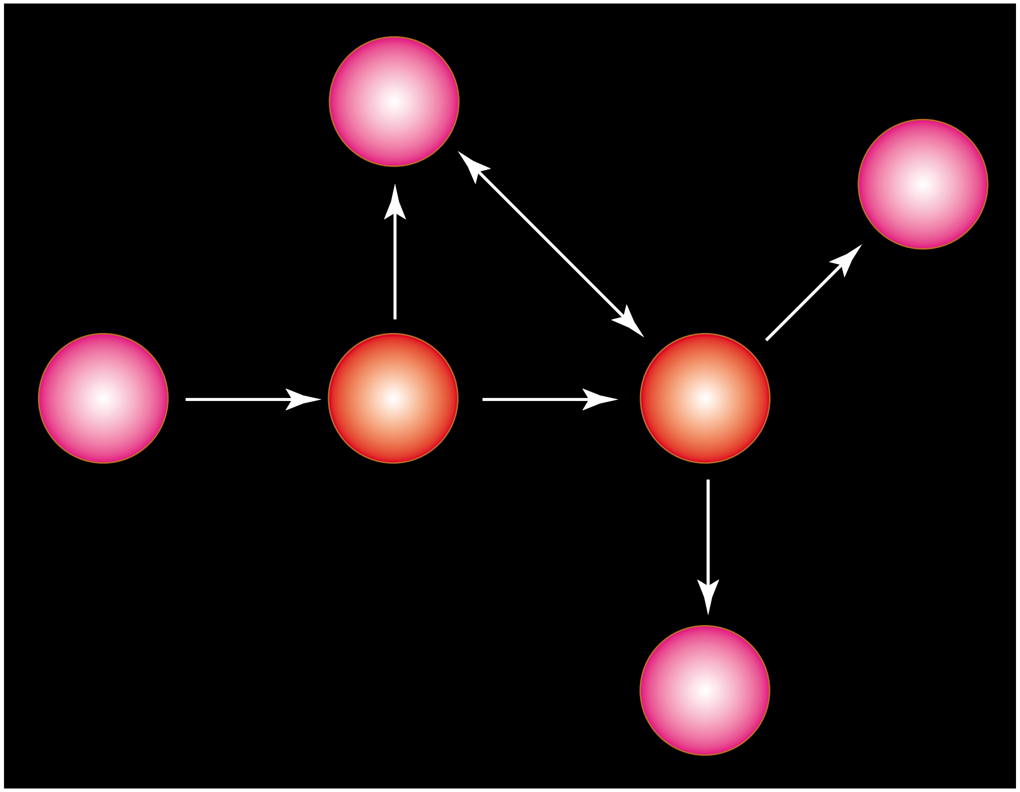
Compromise

1. Do direct query matching



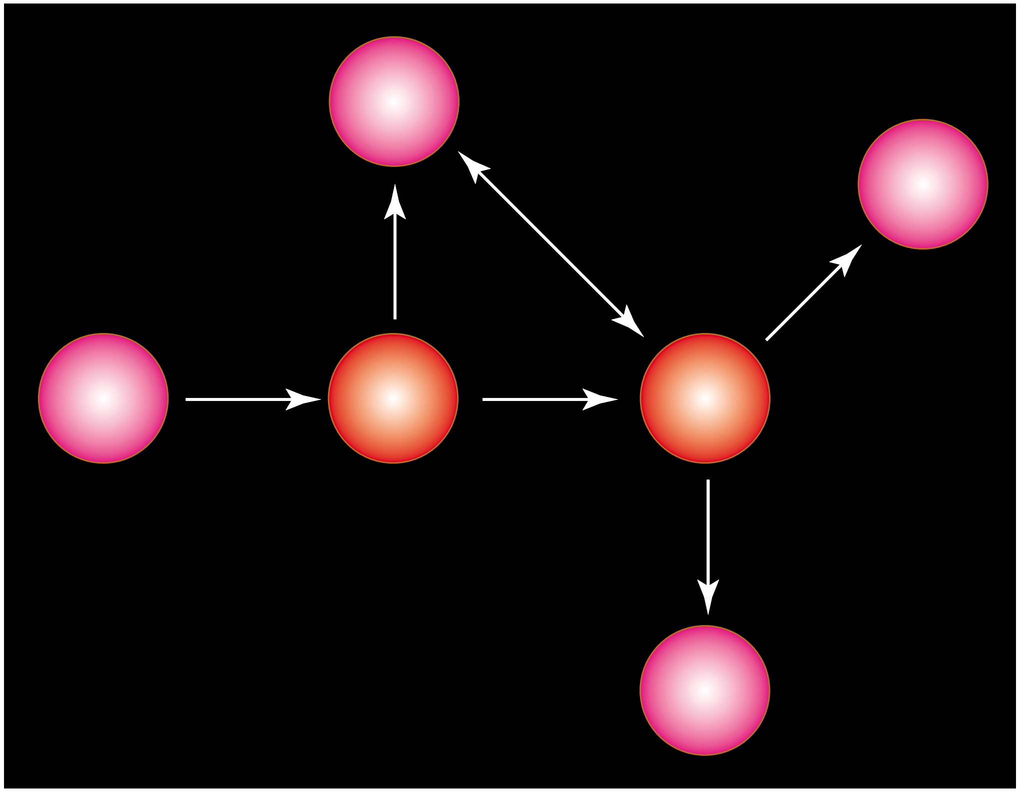
Compromise

1. Do direct query matching
2. Build neighborhood graph



Compromise

1. Do direct query matching
2. Build neighborhood graph



3. Compute authority & hub scores for just the neighborhood

Pros & Cons

Advantages

- Returns satisfactory results
  - Client gets both authority & hub scores

Pros & Cons

Advantages

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

Pros & Cons

Advantages

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

Disadvantages

- Too much has to happen while client is waiting

Pros & Cons

Advantages

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

Disadvantages

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

Pros & Cons

Advantages

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

Disadvantages

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

Pros & Cons

Advantages

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

Disadvantages

- Too much has to happen while client is waiting
  - Custom built neighborhood graph needed for each query
  - Two eigenvector computations needed for each query
- Scores can be manipulated by creating artificial hubs

HITS Applied

TEOMA™

Ask .com™


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

The Definition

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

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

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

PageRank

The Definition

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

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

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

Successive Refinement

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

PageRank

The Definition

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

$\mathcal{B}_P$ = {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

The Definition

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

$\mathcal{B}_P$ = {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

The Definition

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

$\mathcal{B}_P$ = {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

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

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

$$\ddots$$

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

In Matrix Notation

After Step k

— $\boldsymbol{\pi}_k^T = [r_k(P_{\mathbf{1}}),\ r_k(P_{\mathbf{2}}),\ \cdots,\ r_k(P_n)]$

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 $h_{ij} = \begin{cases} \mathbf{1}/|P_i| & \text{if } i \to j \\ \mathbf{0} & \text{otherwise} \end{cases}$

In Matrix Notation

After Step k

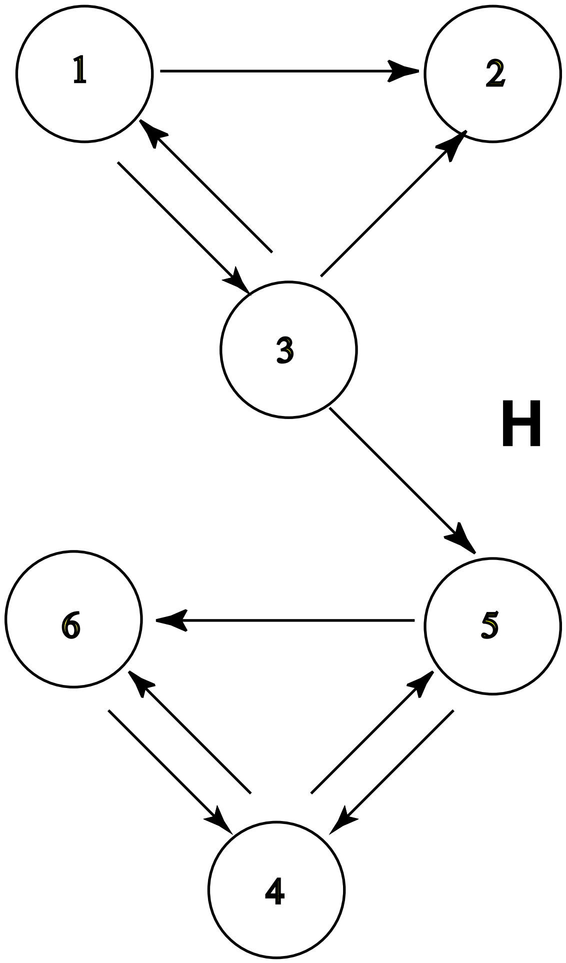
— $\boldsymbol{\pi}_k^T = [r_k(P_1), r_k(P_2), \cdots, r_k(P_n)]$

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

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

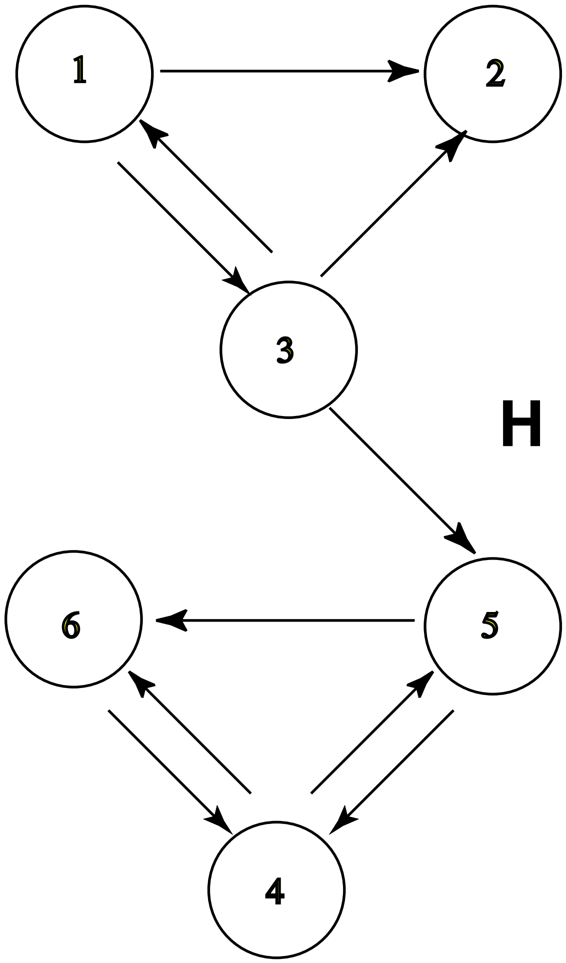
Provided that the limit exists

Tiny Web



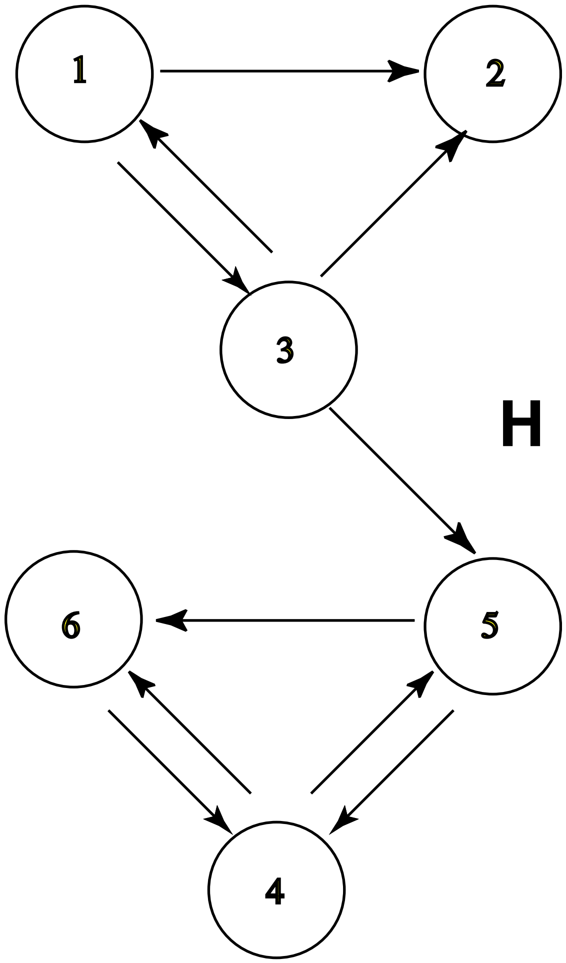
$$\mathbf{H} = \begin{array}{cc} & \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{array}{c} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} & \left(\begin{array}{cccccc} \; & \; & \; & \; & \; & \; \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{array} \right) \end{array}$$

Tiny Web



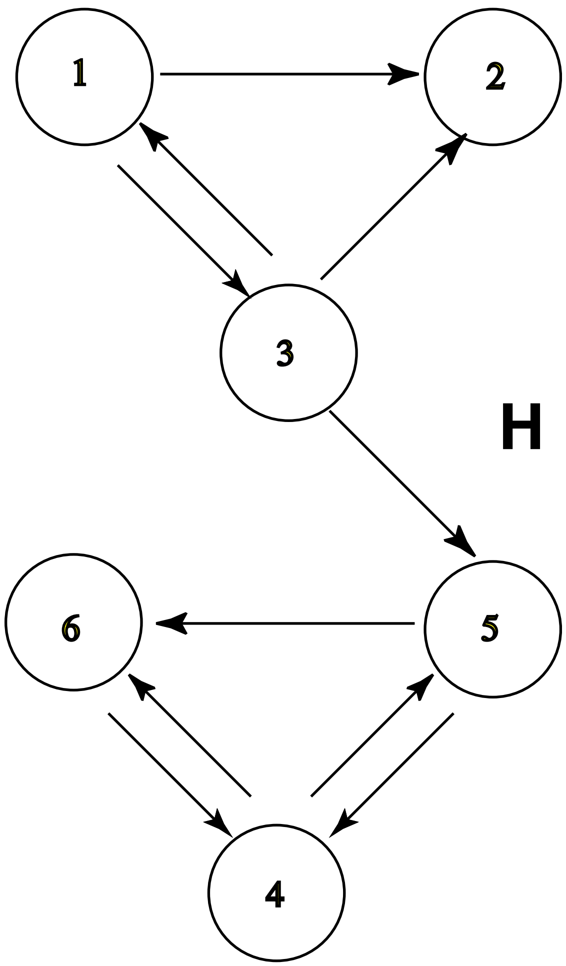
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \left(\begin{array}{cccccc} \mathbf{0} & \mathbf{1/2} & \mathbf{1/2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ & & & & & \\ & &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

Tiny Web



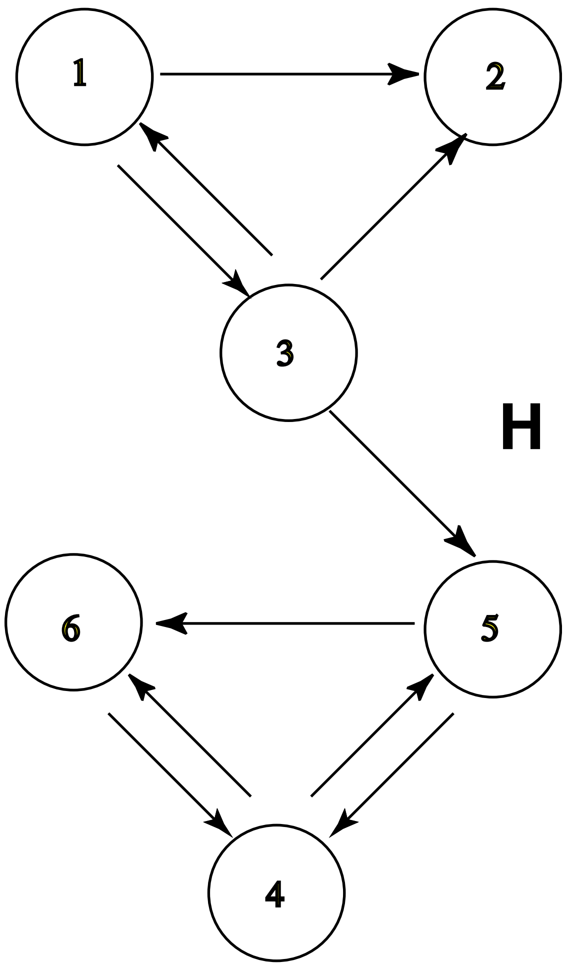
$$\mathbf{H} = \begin{array}{c c} & \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{array}{c} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} & \left(\begin{array}{cccccc} \mathbf{0} & \mathbf{1/2} & \mathbf{1/2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} &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

Tiny Web



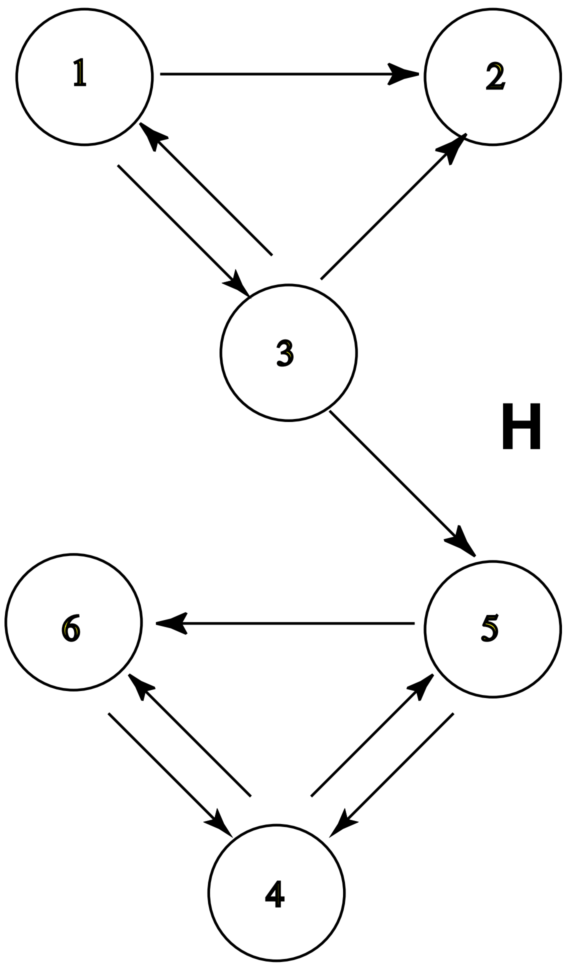
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} \mathbf{0} & \mathbf{1/2} & \mathbf{1/2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}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

Tiny Web



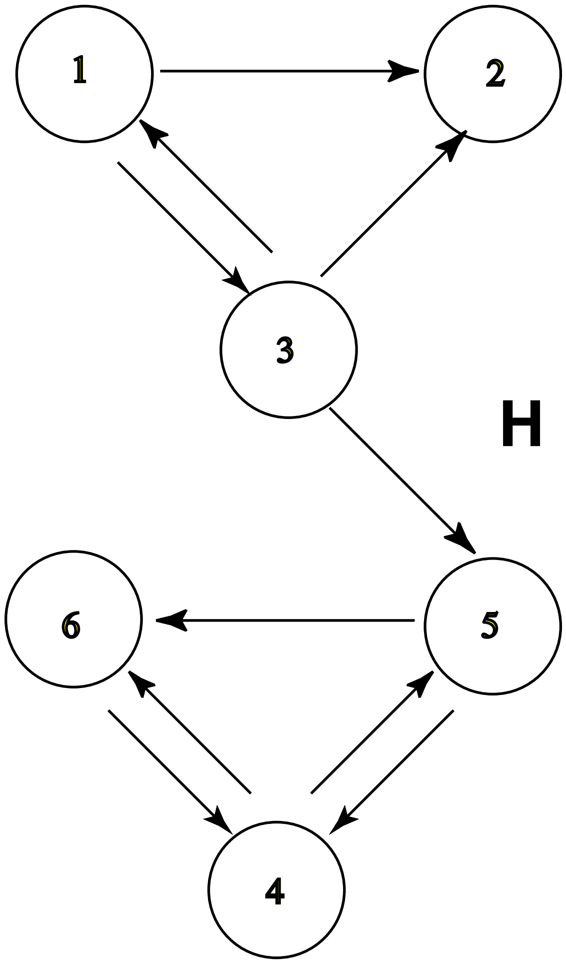
$$
\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array}
\begin{array}{c}
\begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\
\begin{pmatrix}
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2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

Tiny Web



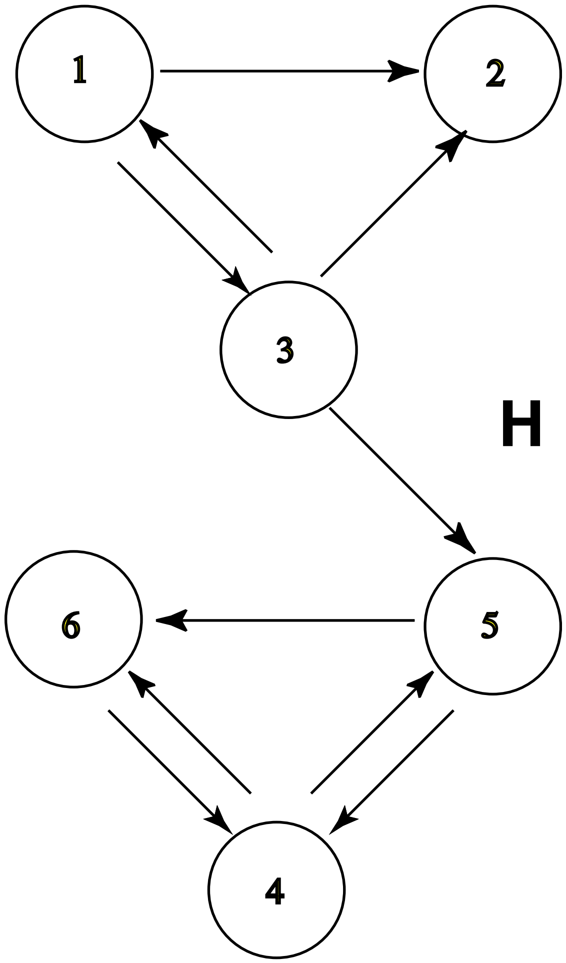
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} \mathbf{0} & \mathbf{1/2} & \mathbf{1/2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}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

Tiny Web



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

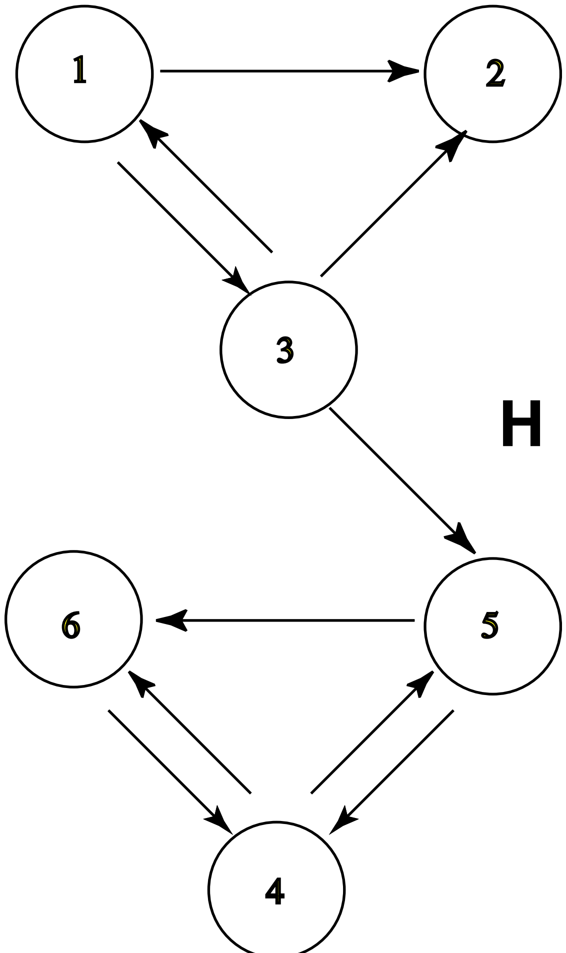
Tiny Web



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

- A random walk on the Web Graph

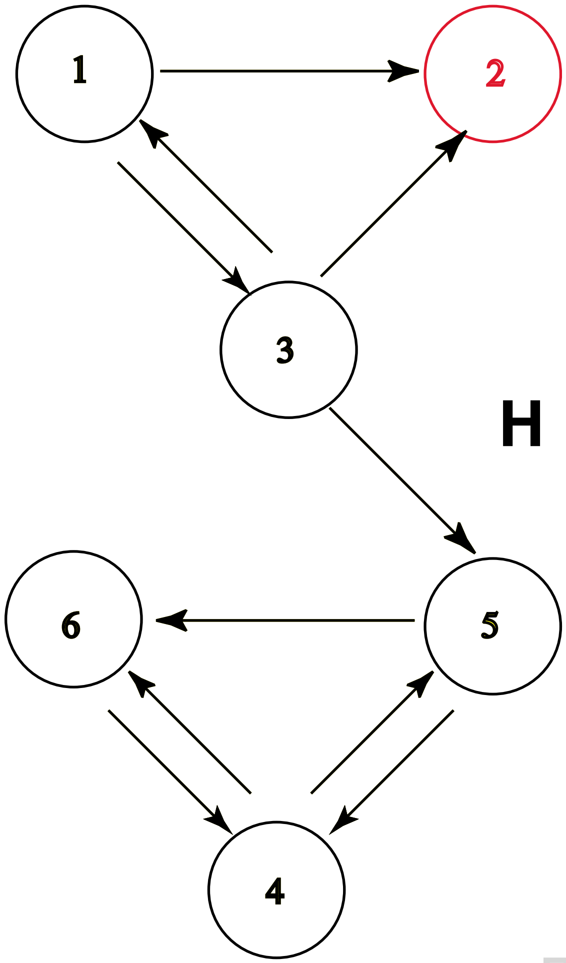
Tiny Web



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

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

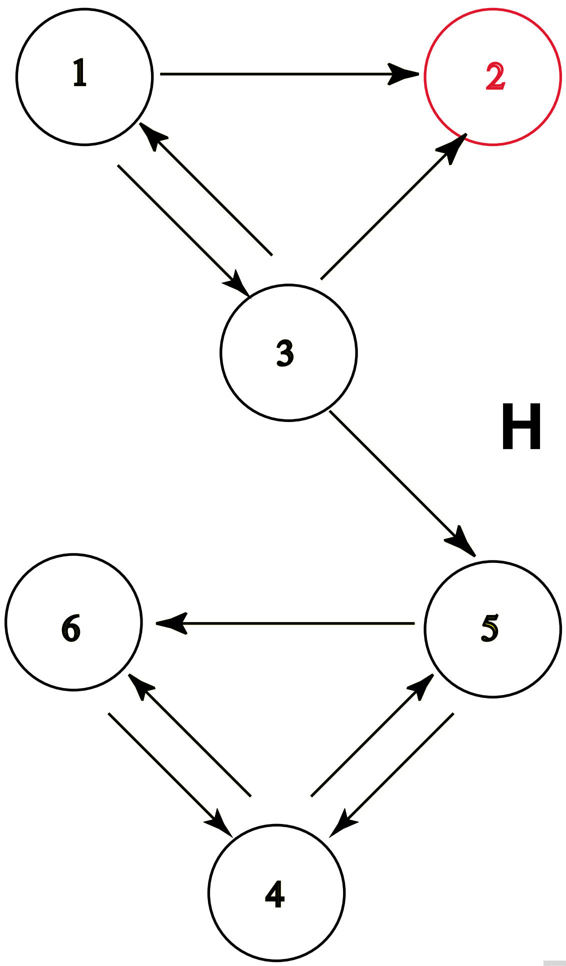
Tiny Web



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

- ▷ 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 $\dfrac{\mathbf{e}^T}{n} = \left(\dfrac{1}{n}, \dfrac{1}{n}, \ldots, \dfrac{1}{n}\right)$

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

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

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

- $\mathbf{S} = \mathbf{H} + \frac{\mathbf{a}\,\mathbf{e}^T}{6}$ is now row stochastic $\Longrightarrow$ $\rho(\mathbf{S}) = 1$
- 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 Is Not Strongly Connected

Nasty Problem

The Web Is Not Strongly Connected

- S is reducible

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

Nasty Problem

The Web Is Not Strongly Connected

- S is reducible

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

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

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

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

Irreducibility Is Not Enough

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

- The powers $\mathbf{S}^k$ fail to converge
- $\boldsymbol{\pi}_{k+1}^T = \boldsymbol{\pi}_k^T\mathbf{S}$ fails to convergence

Irreducibility Is Not Enough

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

— 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 $(P_i \rightarrow P_j \rightarrow P_i)$

- 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 $\quad \boldsymbol{\pi}^T$ = 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 $\qquad \pi^T$ = 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 $\qquad \pi^T$ = 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 $\qquad \pi^T$ = 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 $\qquad \pi^T$ = 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 build-up. (Column 4 and Pages A4 and A6)

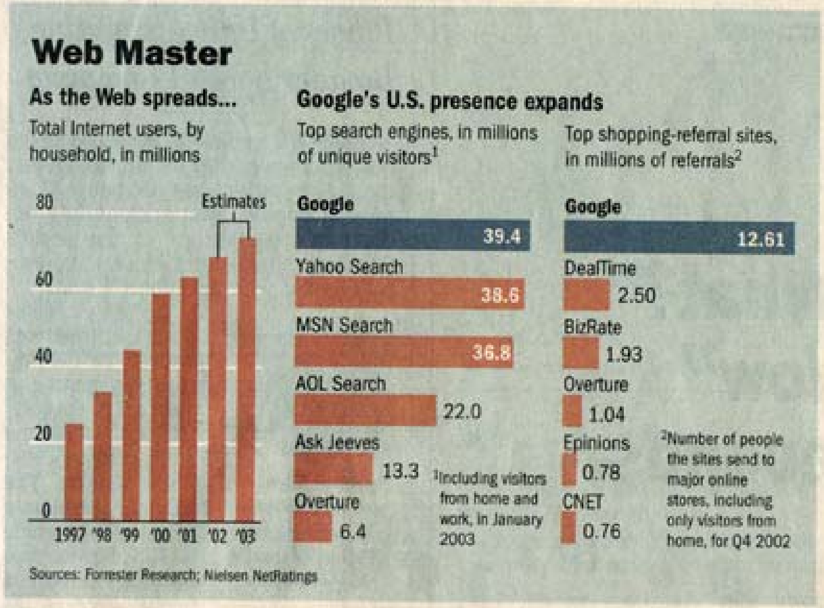
Turkey's parliament debated legislation to let the U.S. deploy 62,000 to open a northern front. Kurdish soldiers lined roads in a show of force as U.S. officials traveled into Iraq's north for an opposition conference.

* * *

■ **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 worked...

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. ...

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 Web-site 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 ar analysts, the company's ch chief executive, Bob Nar Home Depot is prepared t dissatisfied customers and competitive challenge from val with remodeled stores, i ventory and improved custon

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 compara periods, the company said qu increased 5% and net income

Same-store sales, or sale 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 uni

Mr. Agnelli, the 68-year-o Fiat patriarch Gianni Agnel 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

Jason Raush

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 re-ordered 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 Linear Algebra

- **That Is Changing The World**