Google's PageRank and Beyond:

The Science of Search Engine Rankings

Amy Langville

langvillea@cofc.edu

Department of Mathematics
College of Charleston
Charleston, SC

SMU 11/14/07

Google's PageRank and Beyond

THE SCIENCE OF SEARCH ENGINE RANKINGS

AMY N. LANGVILLE and CARL D. MEYER

Outline

- Introduction to Information Retrieval
- Elements of a Search Engine
- Link Analysis
- Current Issues in Web Search

Short History of IR

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

B. C.	cave paintings
7-8th cent. A.D.	Beowulf
12th cent. A.D.	invention of paper, monks in scriptoriums
1450	Gutenberg's printing press
1700s	Franklin's public libraries
1872	Dewey's decimal system
	Card catalog
1940s-1950s	Computer
1960s	Salton's SMART system
1989	Berner-Lee's WWW

System for the Mechanical Analysis and Retrieval of Text

Harvard 1962 – 1965

Cornell 1965 – 1970

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?

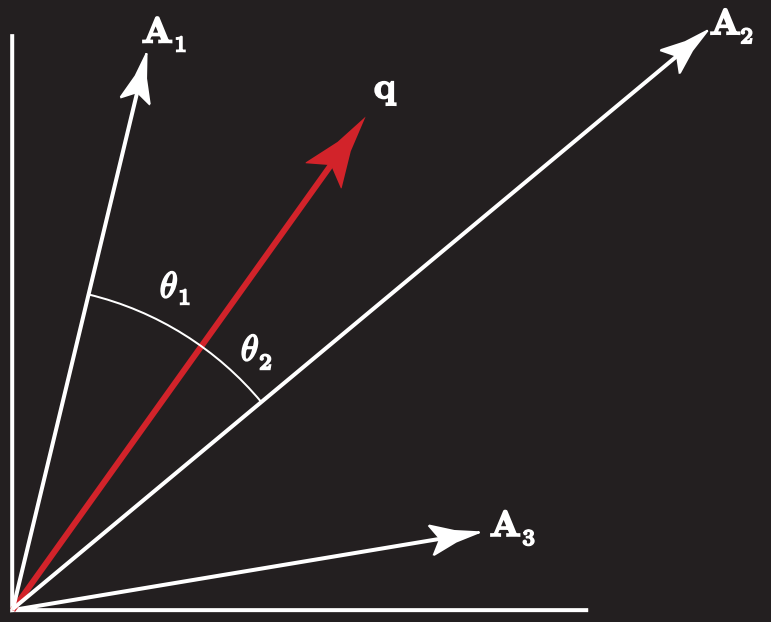
Query Matching

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 $\mathbf{A}_i$?



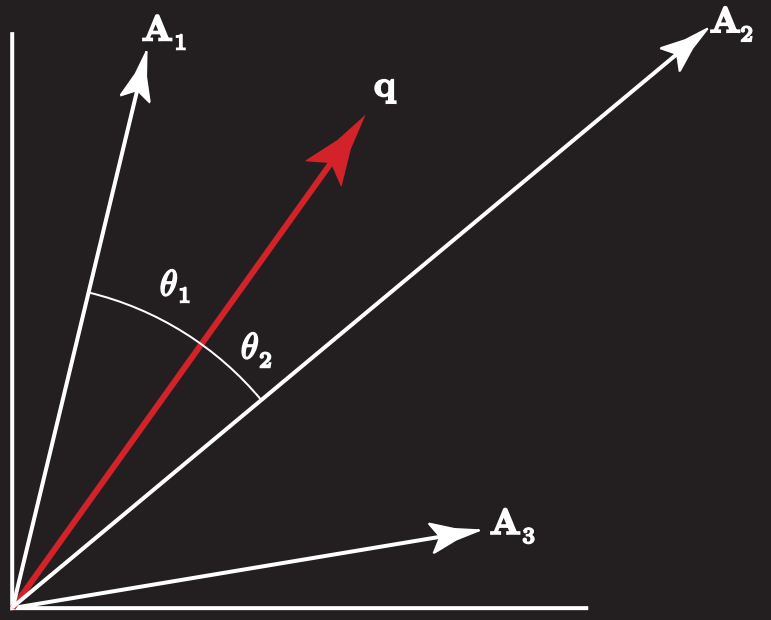
Query Matching

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

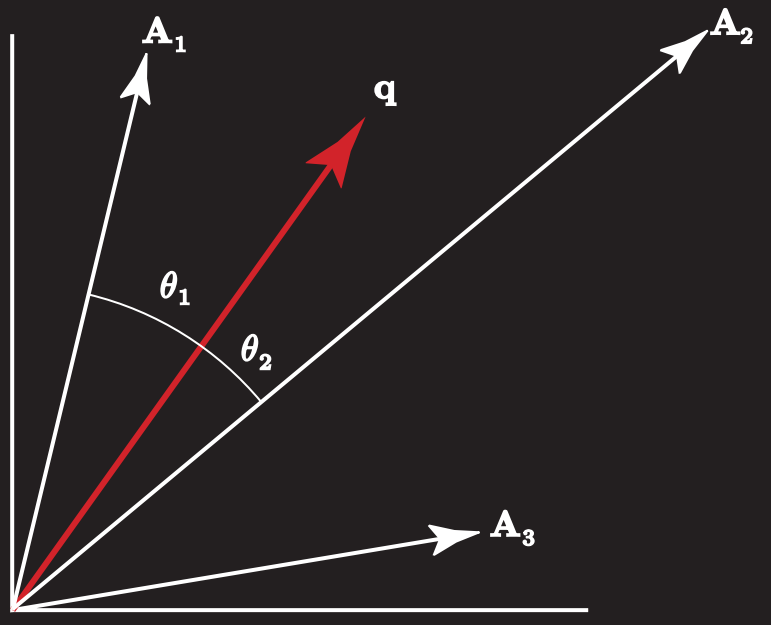
Query Matching

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 $\mathbf{A}_i$?



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

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

Traditional IR

Pros

- Finds hidden connections

Traditional IR

Pros

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

Traditional IR

Pros

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

Traditional IR

Pros

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

Cons

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

Traditional IR

Pros

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

Cons

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

Traditional IR

Pros

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

Cons

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

Traditional IR

Pros

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

Cons

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

Trad. IR applied to Web

the pre-1998 Web Index

⋮

- border patrol: 4; 567; 809; 1103;

⋮

- hezbollah: 9; 12; 339; 942; 15158;

⋮

- global warming: 178; 12980; 445532;

Index

Trad. IR applied to Web

the pre-1998 Web Index

⋮

- border patrol: 4; 567; 809; 1103; . . . (8,700,000 in total)

⋮

- hezbollah: 9; 12; 339; 942; 15158; . . . (15,100,000 in total)

⋮

- global warming: 178; 12980; 445532; . . . (33,200,000 in total)

Trad. IR applied to Web

the pre-1998 Web Index

⋮

- border patrol: 4; 567; 809; 1103; . . . (8,700,000 in total)

⋮

- hezbollah: 9; 12; 339; 942; 15158; . . . (15,100,000 in total)

⋮

- global warming: 178; 12980; 445532; . . . (33,200,000 in total)

too many results per search term

easily spammed

Sentiments about the pre-1998 Web

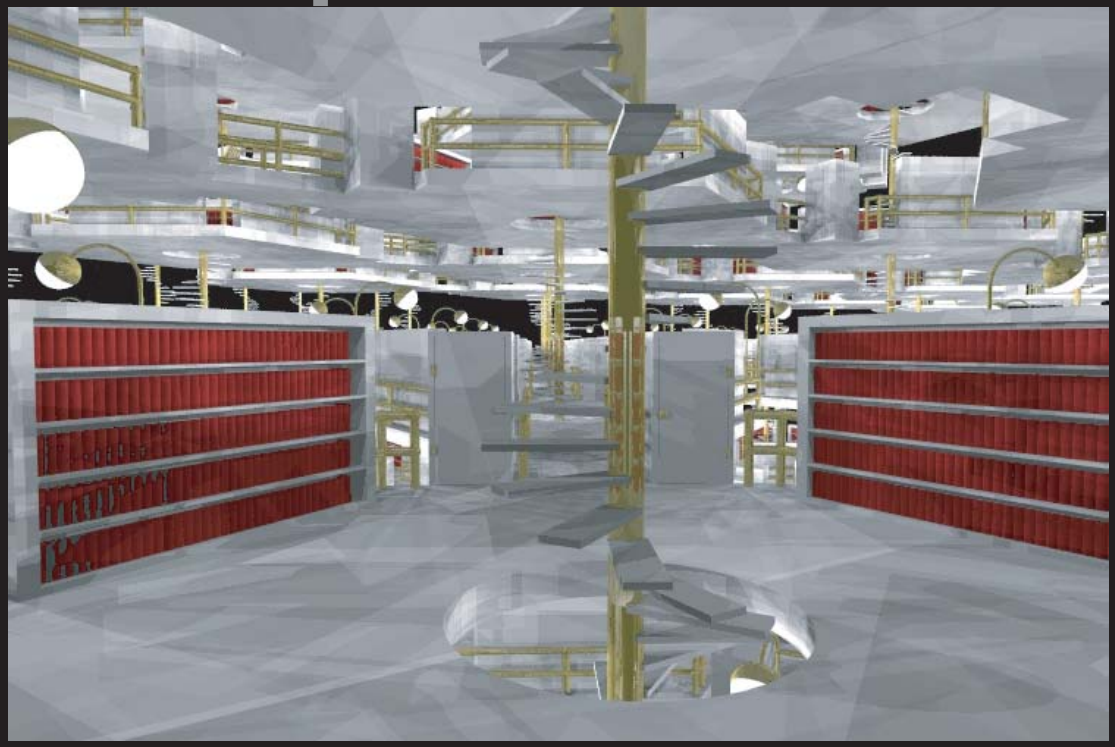
Yahoo

- hierarchies of sites
- organized by humans

Best Search Techniques

- word of mouth
- expert advice

Overall Feeling of Users



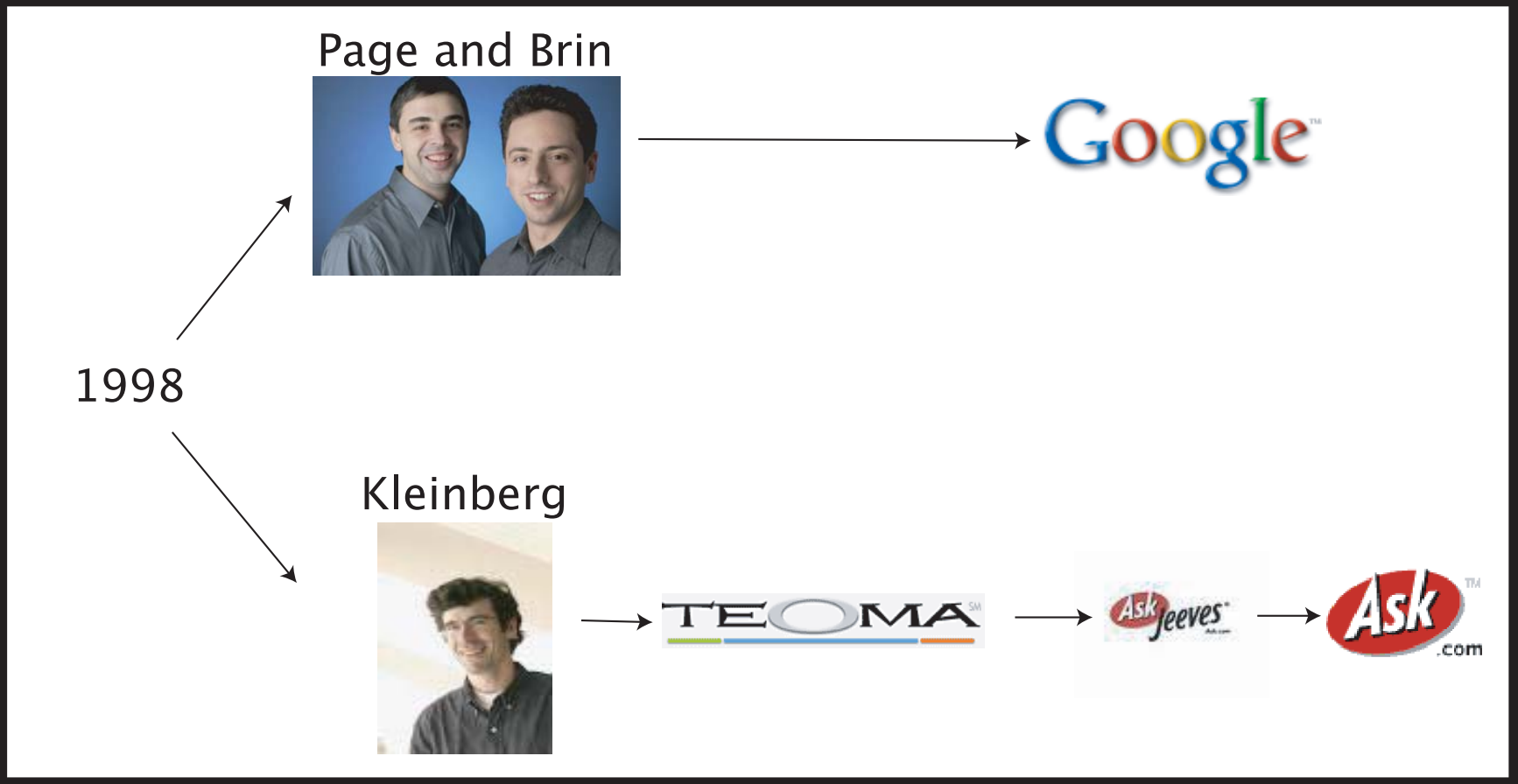
- Jorge Luis Borges' 1941 short story, *The Library of Babel*

When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness. All men felt themselves to be the masters of an intact and secret treasure. There was no personal or world problem whose eloquent solution did not exist in some hexagon.

... As was natural, this inordinate hope was followed by an excessive depression. The certitude that some shelf in some hexagon held precious books and that these precious books were inaccessible, seemed almost intolerable.

1998: enter Link Analysis

- uses *hyperlink* structure to focus the relevant set
- combine traditional IR score with popularity score



1998 ... enter Link Analysis

Change in User Attitudes about Web Search

Today

- “It’s not my homepage, but it might as well be. I use it to ego-surf. I use it to read the news. Anytime I want to find out anything, I use it.” - Matt Groening, creator and executive producer, The Simpsons
- “I can’t imagine life without Google News. Thousands of sources from around the world ensure anyone with an Internet connection can stay informed. The diversity of viewpoints available is staggering.” - Michael Powell, chair, Federal Communications Commission
- “Google is my rapid-response research assistant. On the run-up to a deadline, I may use it to check the spelling of a foreign name, to acquire an image of a particular piece of military hardware, to find the exact quote of a public figure, check a stat, translate a phrase, or research the background of a particular corporation. It’s the Swiss Army knife of information retrieval.” - Garry Trudeau, cartoonist and creator, Doonesbury

Web Information Retrieval

IR before the Web = traditional IR

IR on the Web = web IR

Web Information Retrieval

IR before the Web = traditional IR

IR on the Web = web IR

How is the Web different from other document collections?

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web
- It's dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web
- It's dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year
- It's self-organized.
  - no standards, review process, formats
  - errors, falsehoods, link rot, and spammers!

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web
- It's dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year
- It's self-organized.
  - no standards, review process, formats
  - errors, falsehoods, link rot, and spammers!

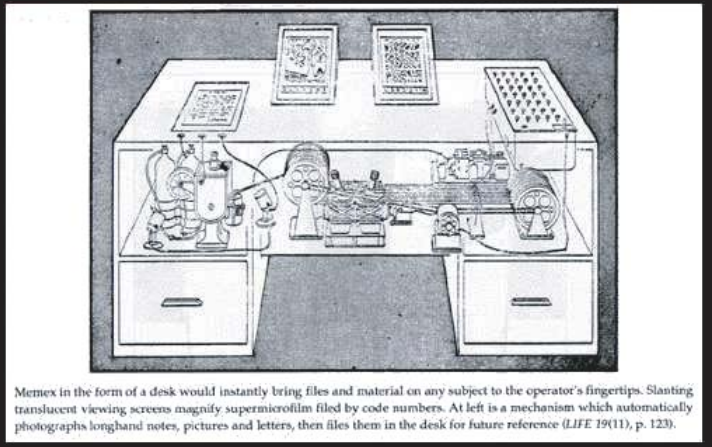
A Herculean Task!

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = **web IR**

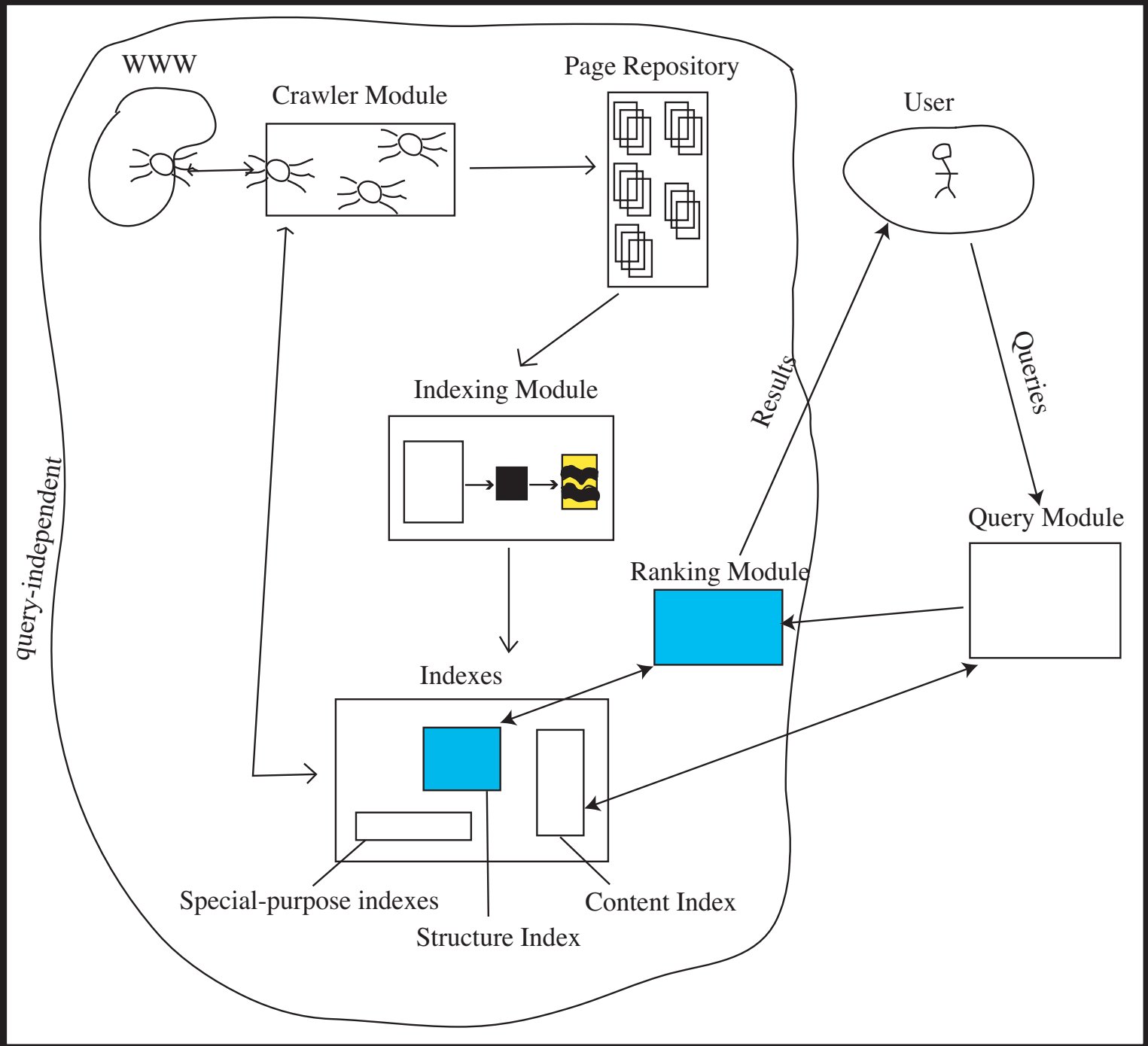
How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, each about 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web
- It's dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year
- It's self-organized.
  - no standards, review process, formats
  - errors, falsehoods, link rot, and spammers!
- Ah, but it's *hyperlinked* !
  - Vannevar Bush's 1945 memex

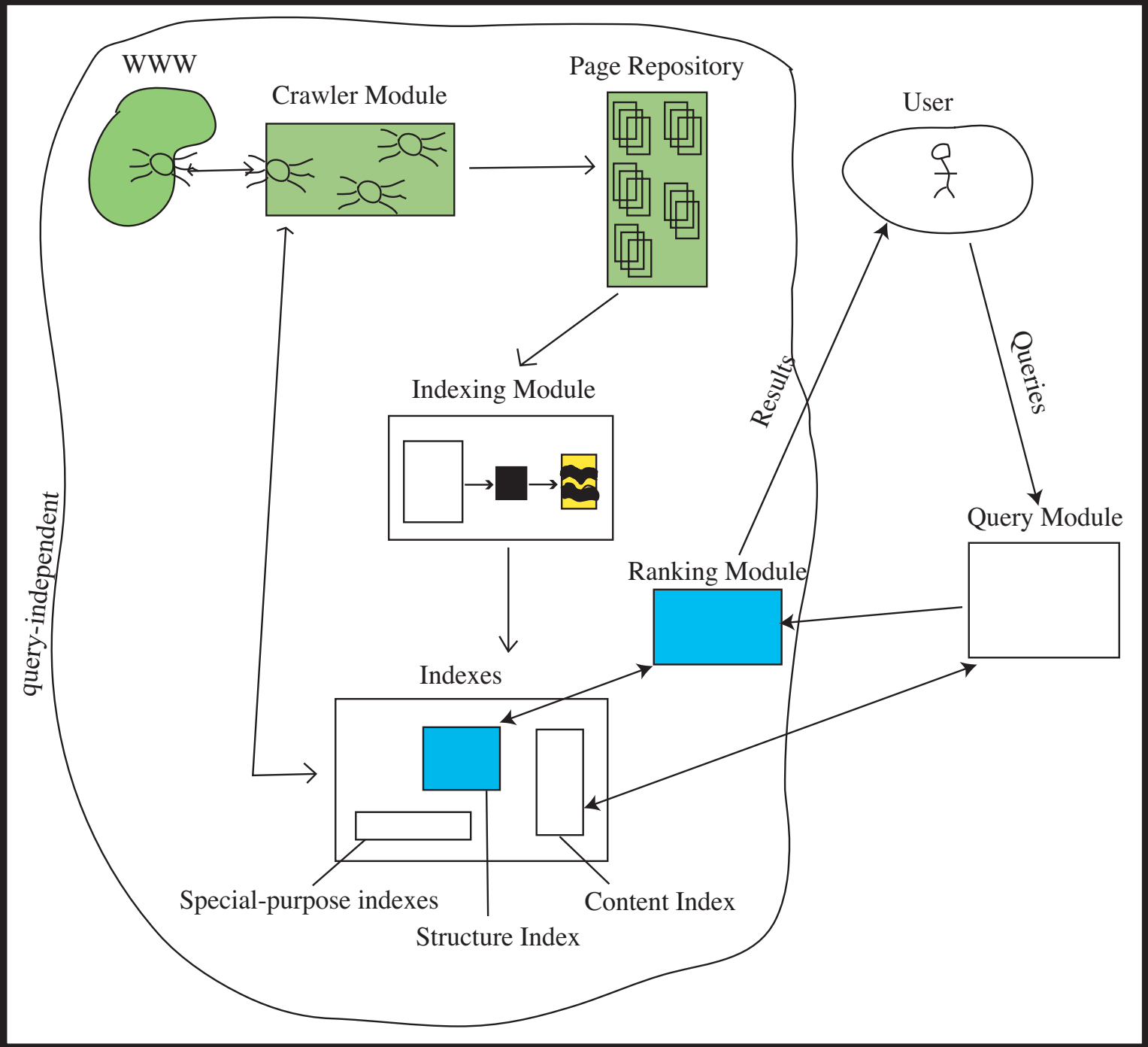


Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE* 19(11), p. 123).

Elements of a Web Search Engine



Elements of a Web Search Engine



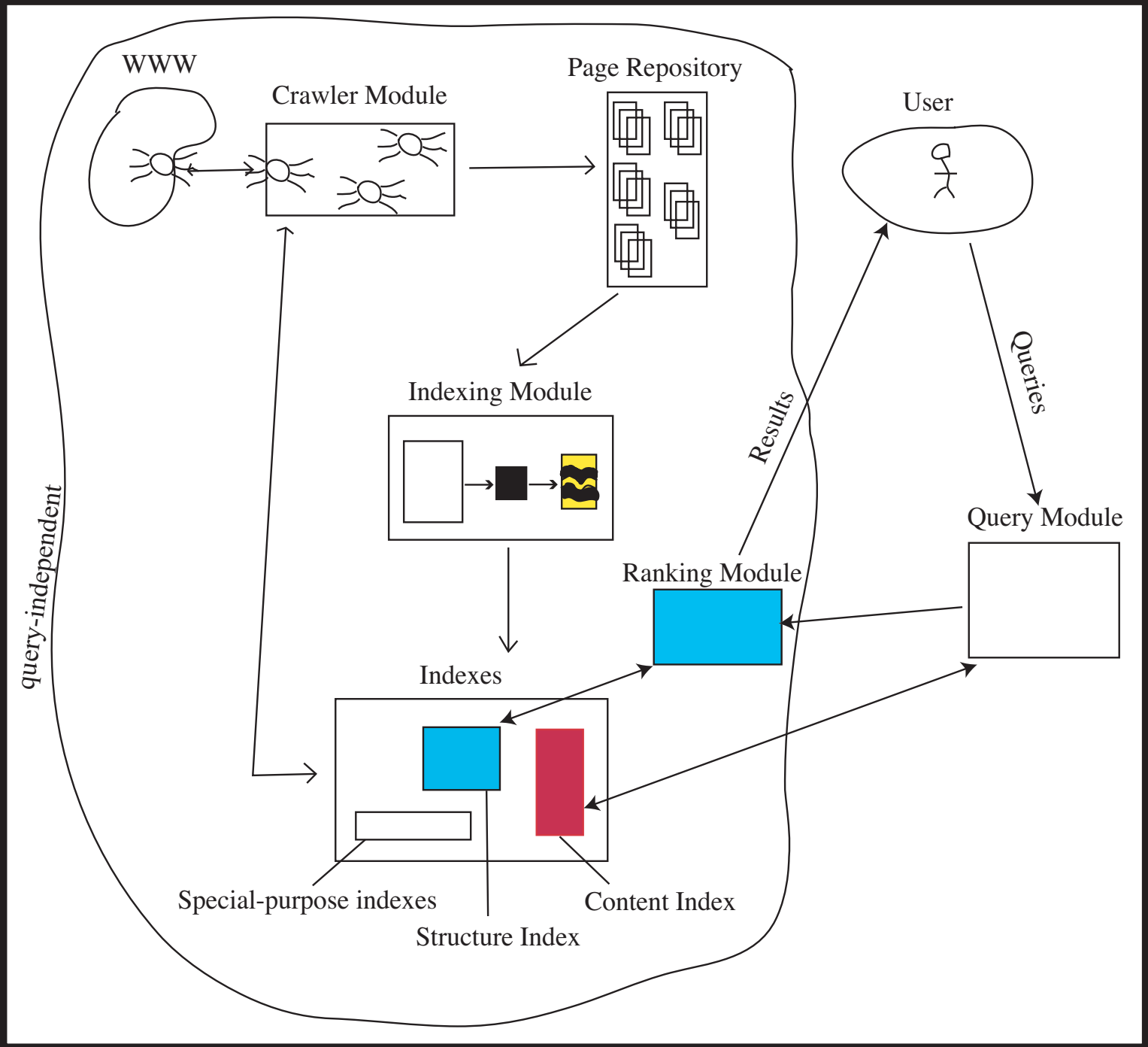
The Ranking Module (generates popularity scores)

- Measure the importance of each page

The Ranking Module (generates popularity scores)

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

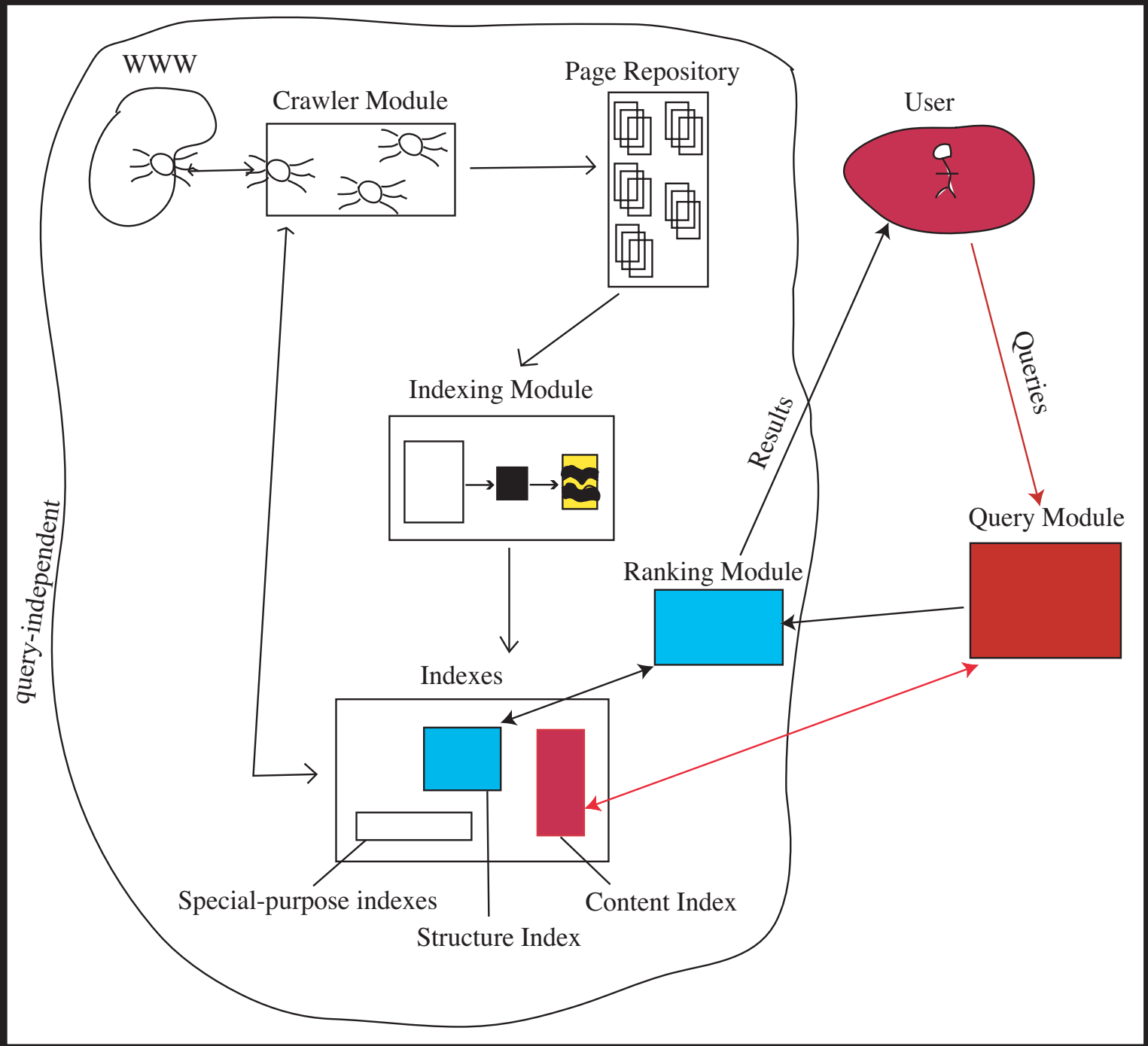
Elements of a Web Search Engine



The Ranking Module (generates popularity scores)

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

Elements of a Web Search Engine



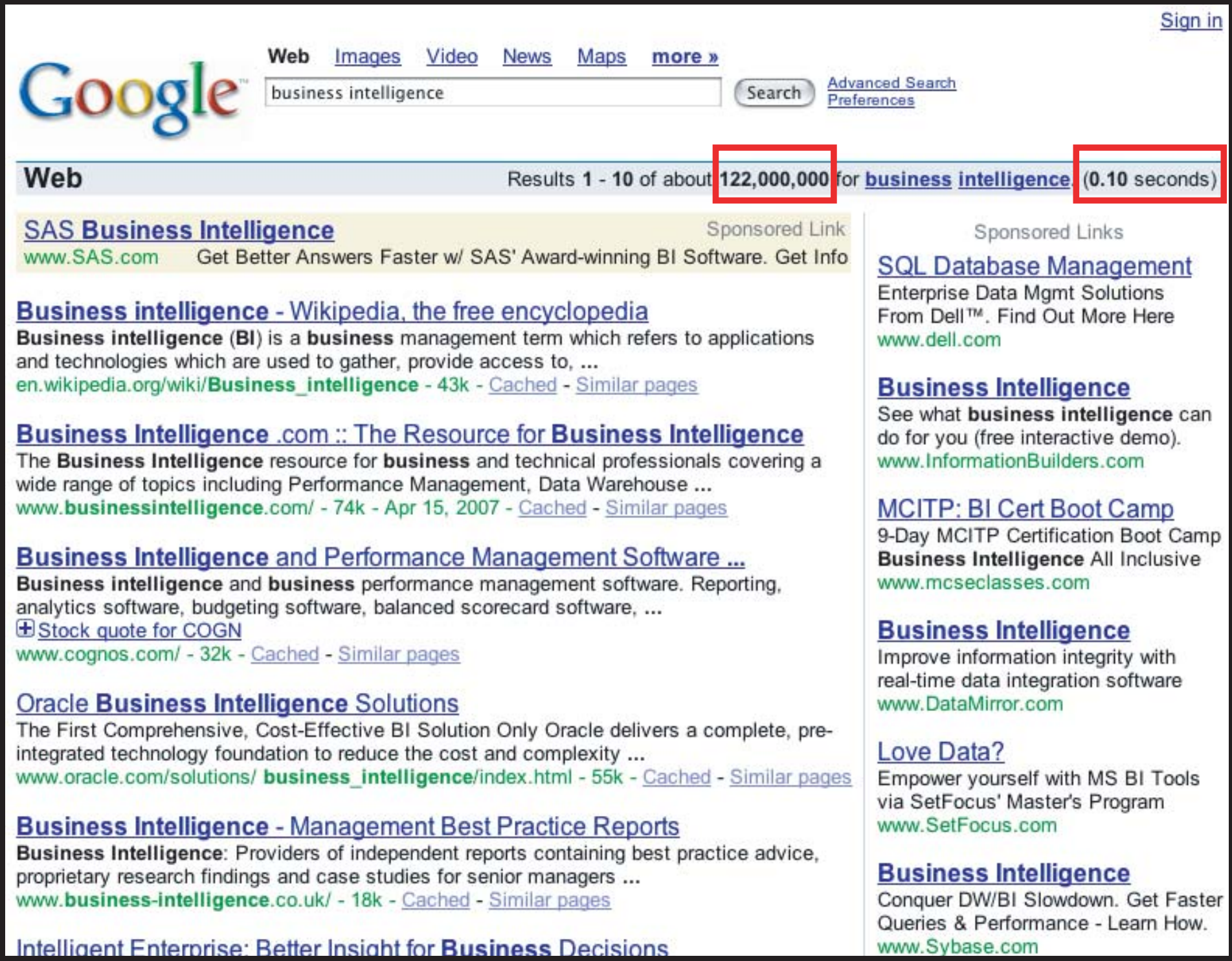
The Ranking Module (generates popularity scores)

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

The Ranking Module (generates popularity scores)

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

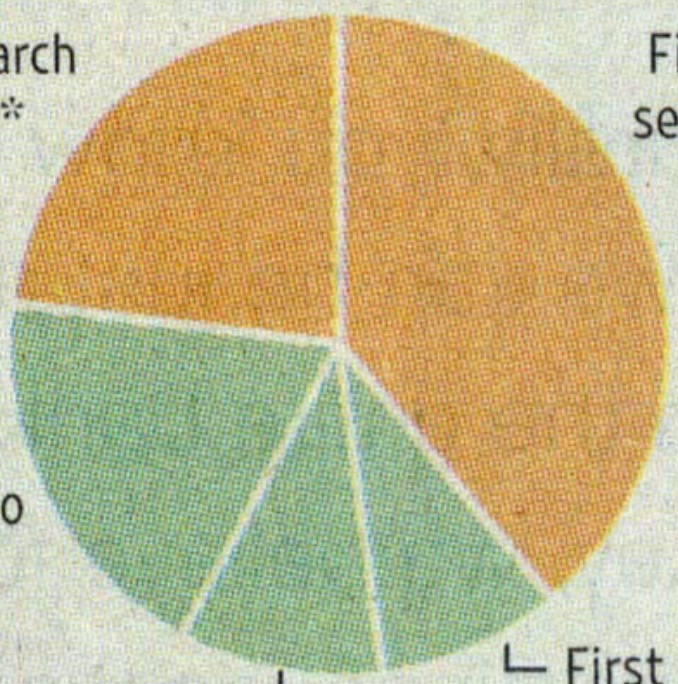
Google's PageRank = Google's $$$$$


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

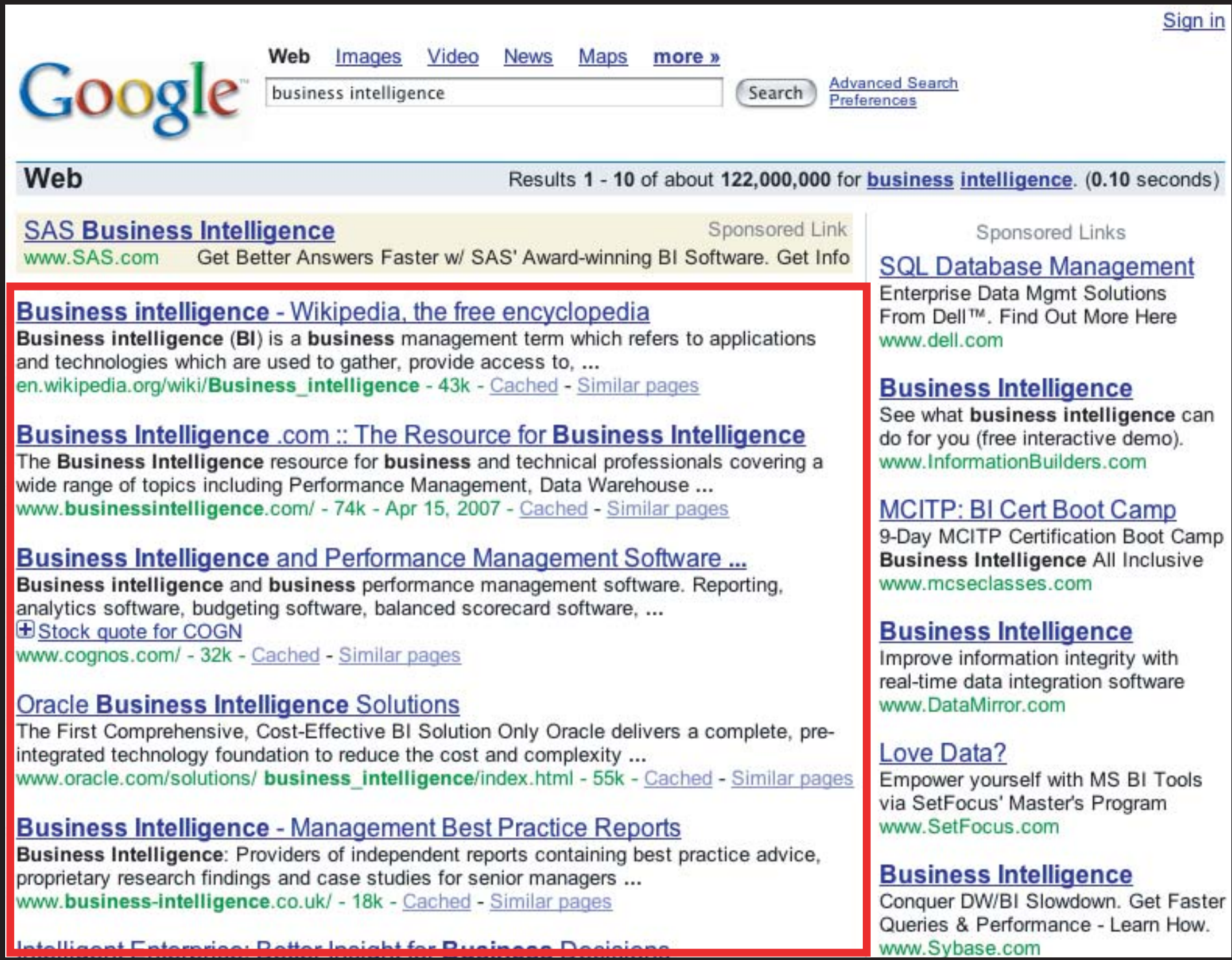
Amount of Internet search results that Web surfers typically scan before selecting one

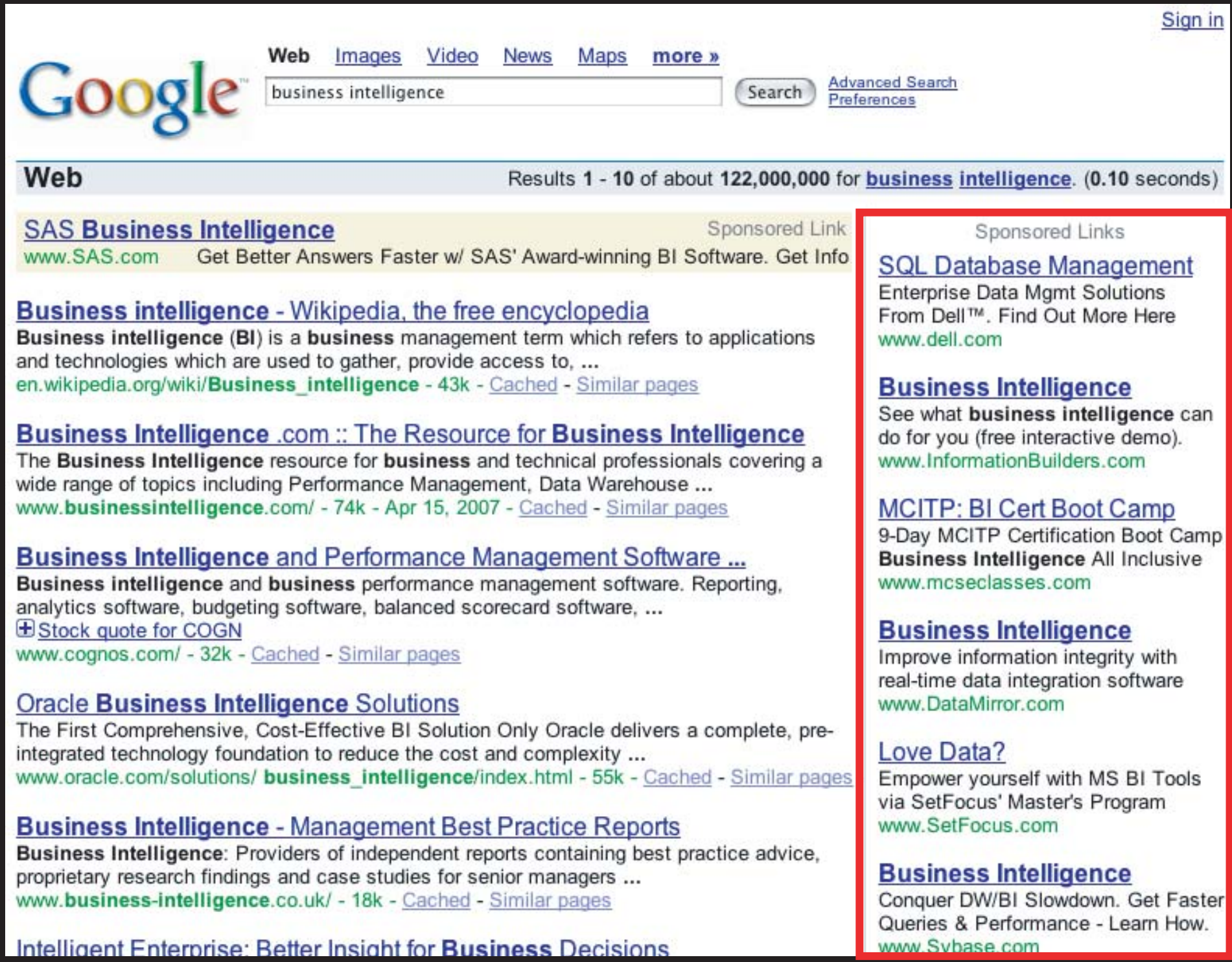


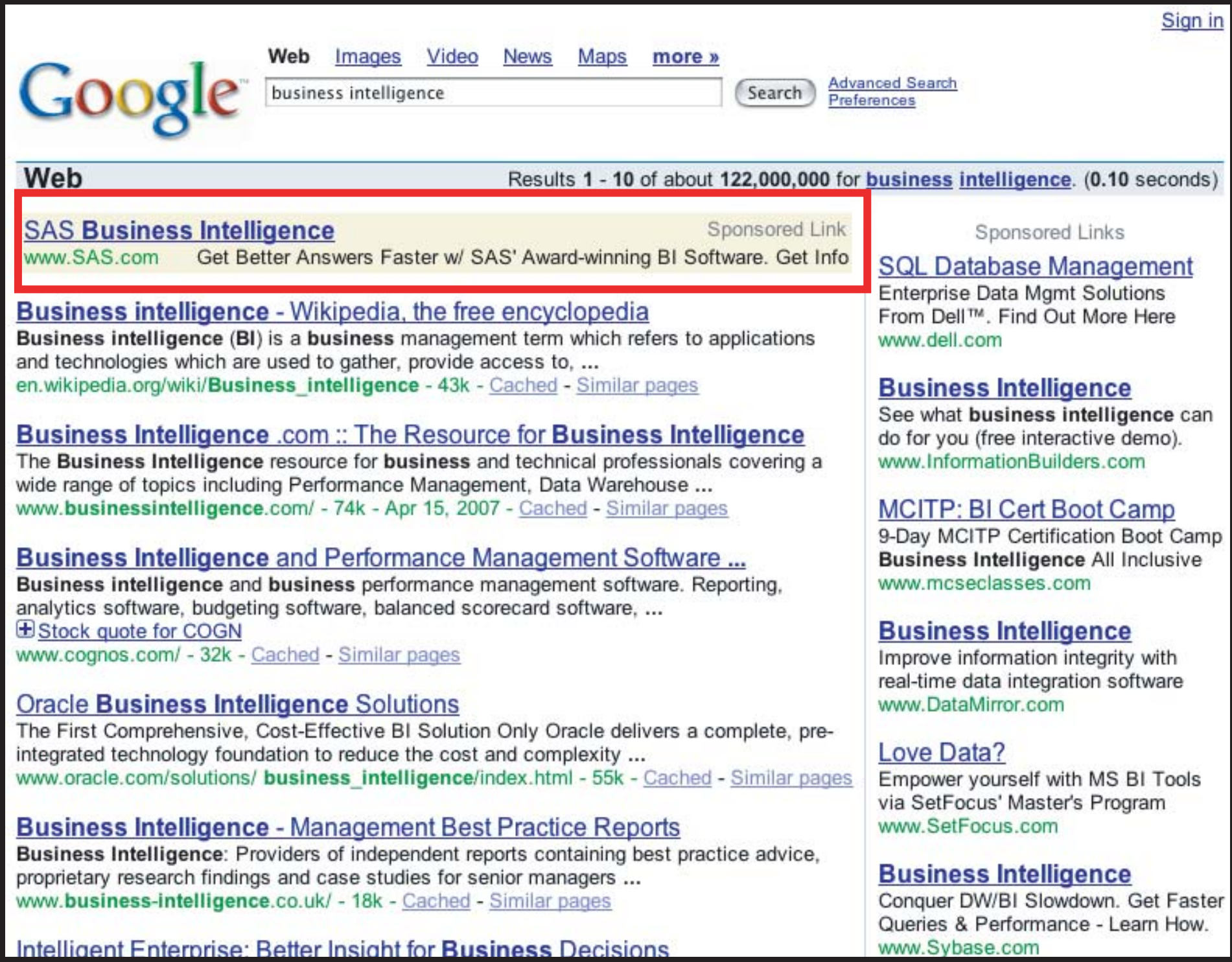
*Top results without reading through the whole page

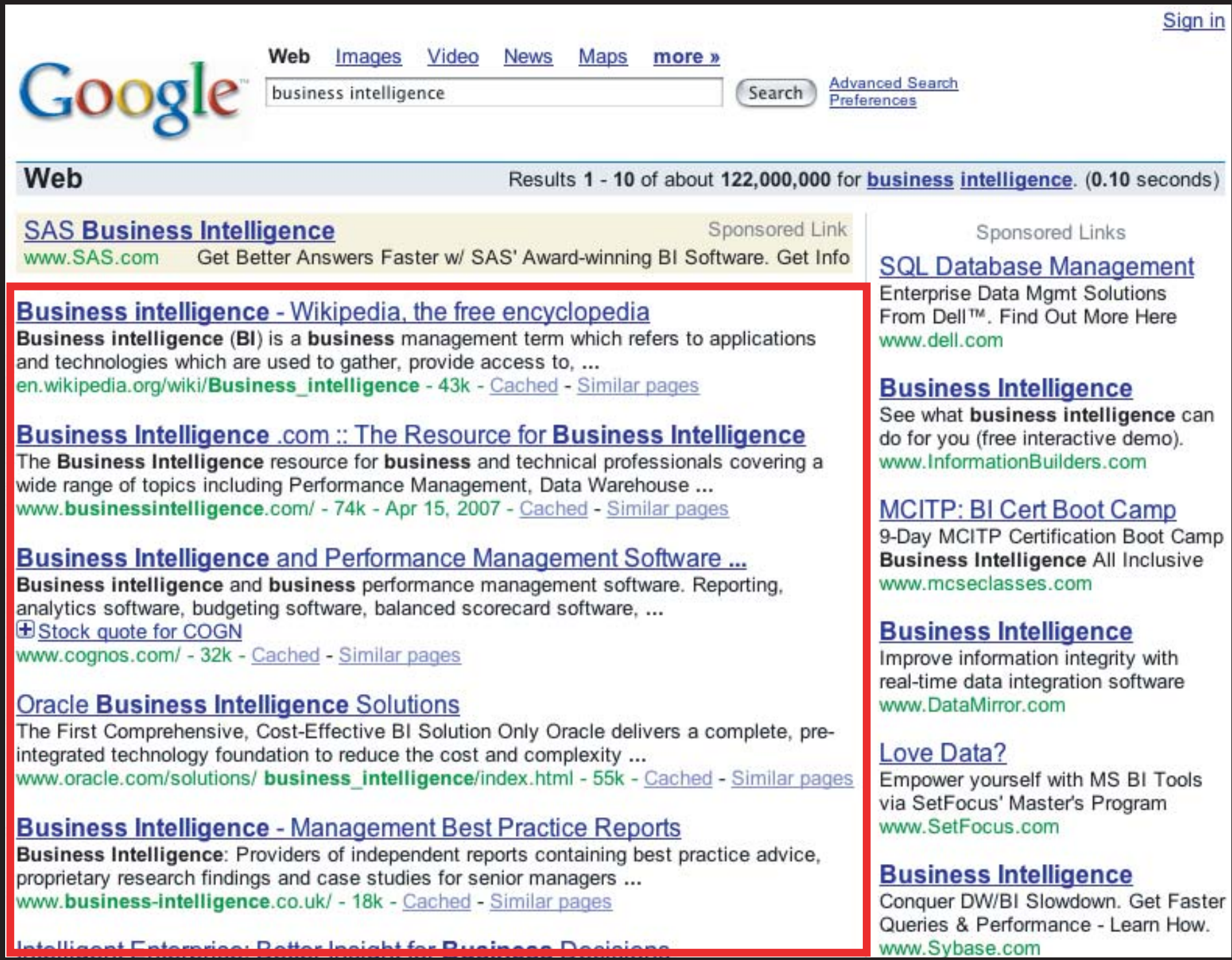
Note: Sample size is 2,369 people

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)
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

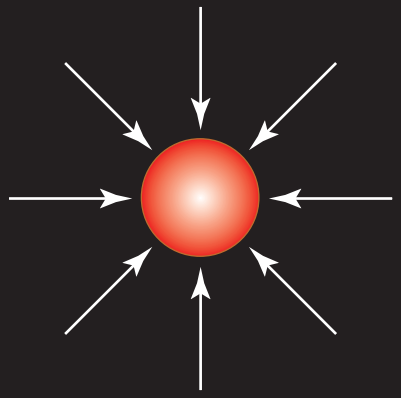

Sign in
Google
Web Images Video News Maps more »
business intelligence
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)
SAS Business Intelligence
Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software.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


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

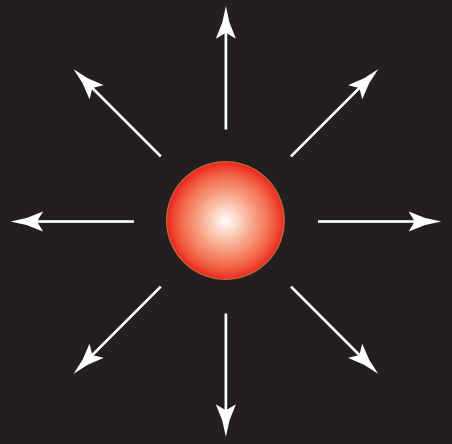

Sign in
Google
Web Images Video News Maps more »
business intelligence
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)
SAS Business Intelligence
Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software.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

How To Measure “Importance”

Landmark Result Paper

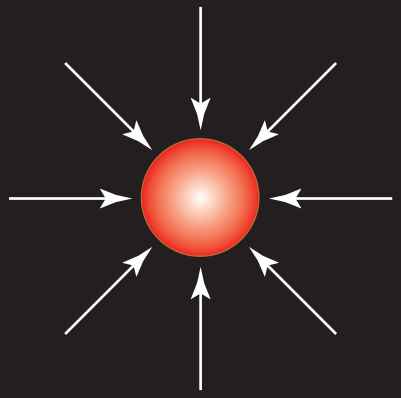


Survey Paper—Big Bib



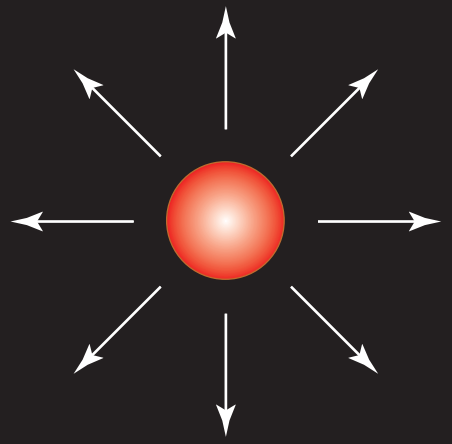
How To Measure “Importance”

Landmark Result Paper



Authorities

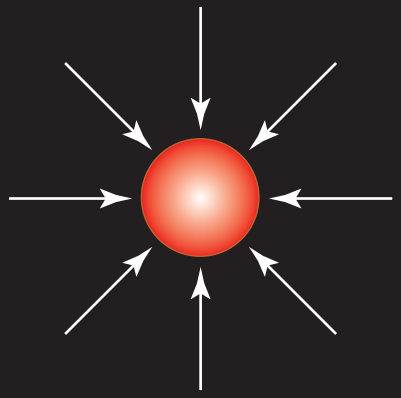
Survey Paper—Big Bib



Hubs

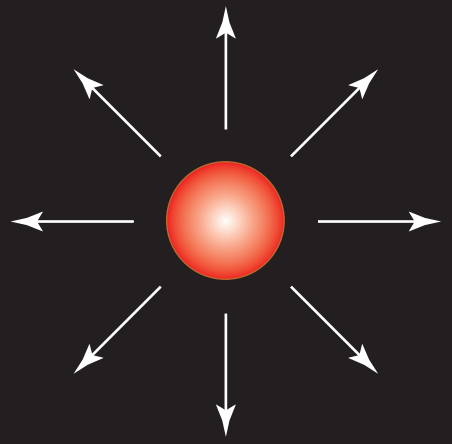
How To Measure “Importance”

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

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

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

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

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

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

Successively Re-refine Authority & Hub Scores

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

Combined Iterations

- $\mathbf{A} = \mathbf{L}^T\mathbf{L}$ (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) $\qquad \mathbf{a}_k = \mathbf{A}\mathbf{a}_{k-1} \to$ 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

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

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)

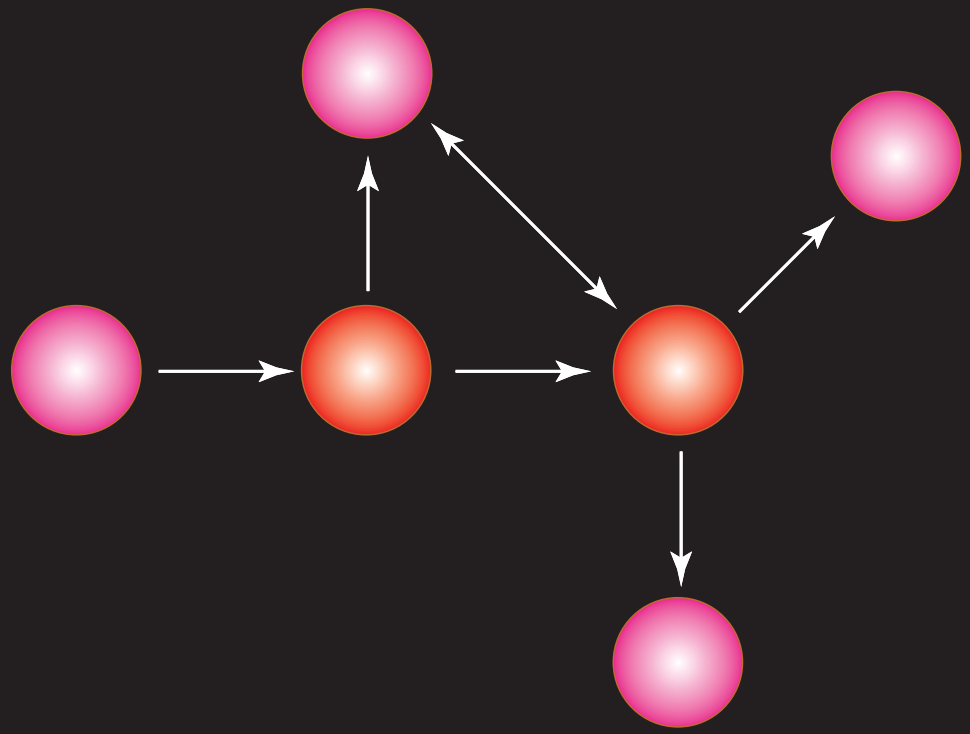
!! **May not be uniquely defined if A or H is reducible** !!

Compromise

1. Do direct query matching

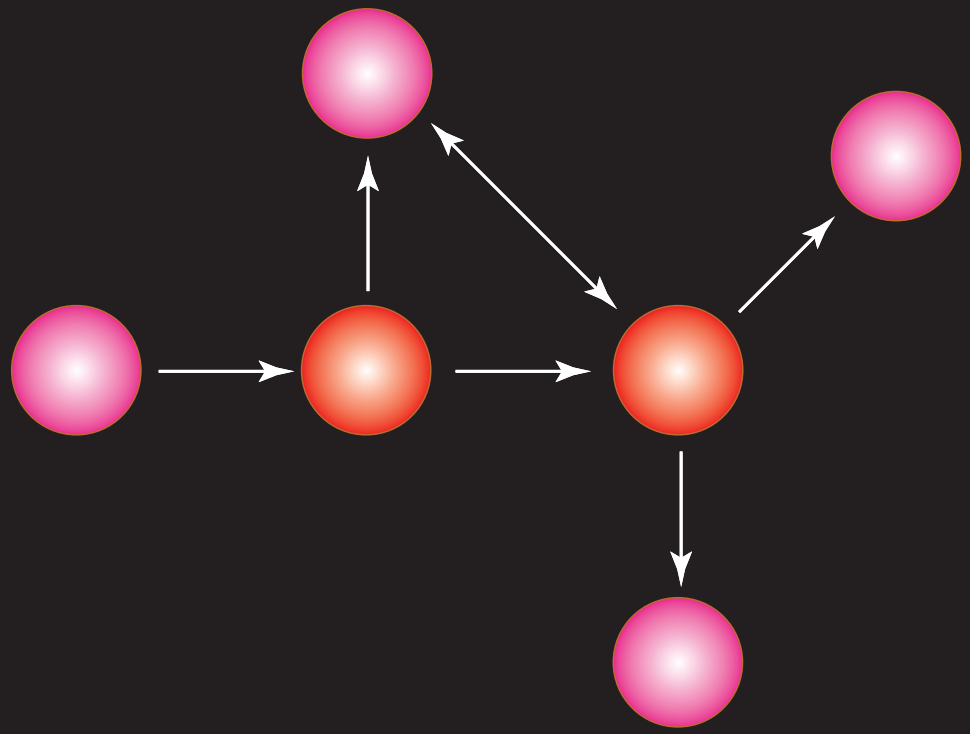
Compromise

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

THE ALGORITHM SEES THE INTERNET THE WAY DMITRY SKLYAROV SEES A POORLY ENCRYPTED DRM FILE.

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.

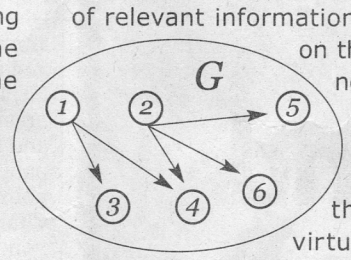
So what's the solution to search? While computer science was trying to coax an answer from its collective hard drive, it was sitting right there in the stacks all along. Who could have guessed that when Eugene Garfield went all bibliometric and devised a system to find out how much a journal mattered 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.

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

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.



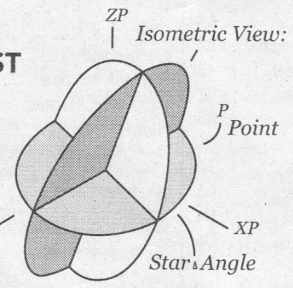
THE ALGORITHM IS RELATIVELY SIMPLE, IF YOU'RE SOME KIND OF SAVANT.

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 $a(p)$ → $a = ATh$ and a non-negative hub weight $h(p)$. → $h = Aa$ 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.

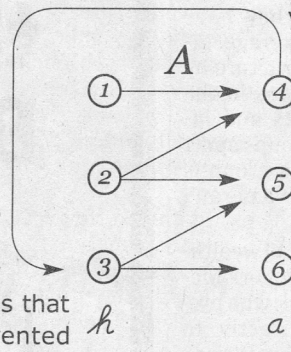
IT'S A GOOD THING ROBERT FROST NEVER WROTE AN ALGORITHM.

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

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

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

$\mathcal{B}_P$ = {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

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

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

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

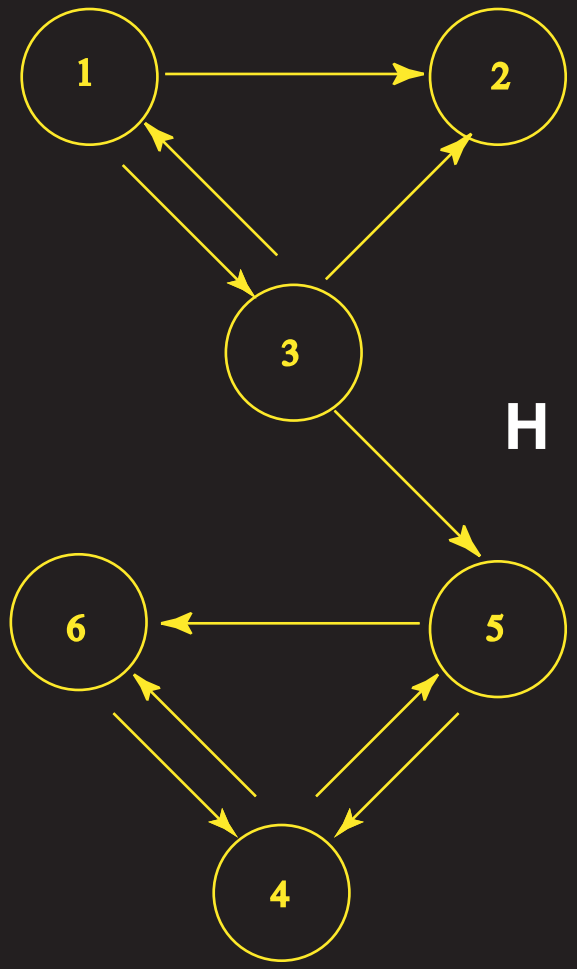
In Matrix Notation

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 $\mathbf{H}$

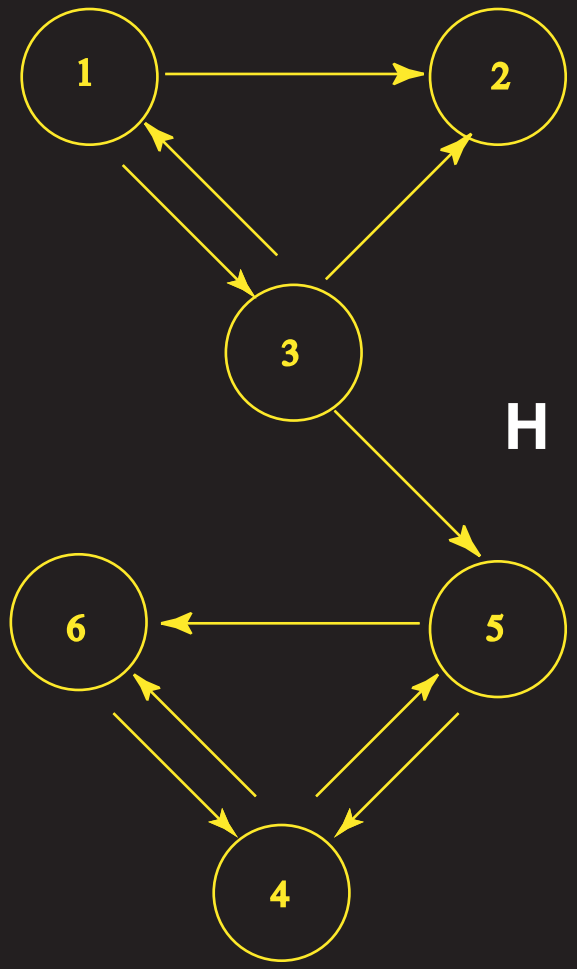
Provided that the limit exists

Tiny Web



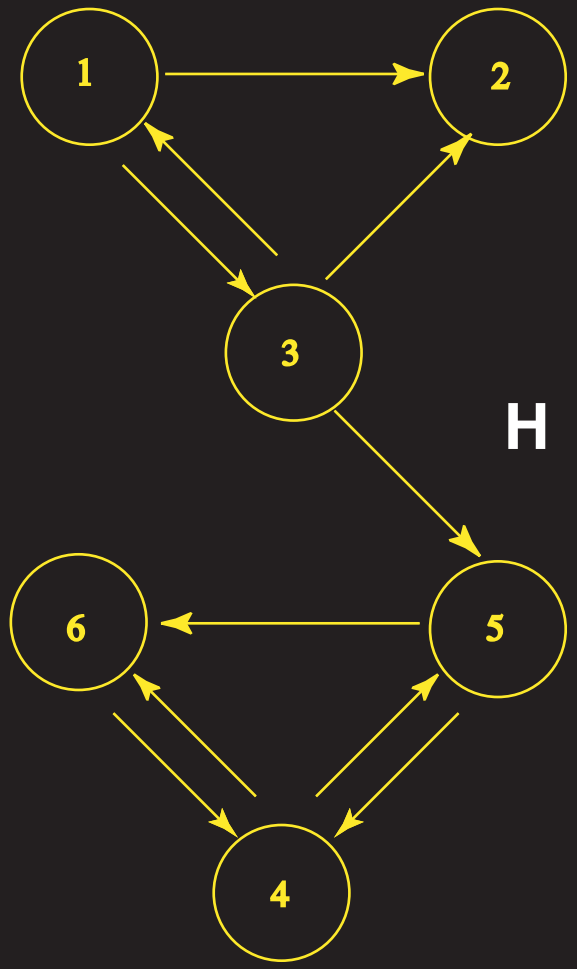
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



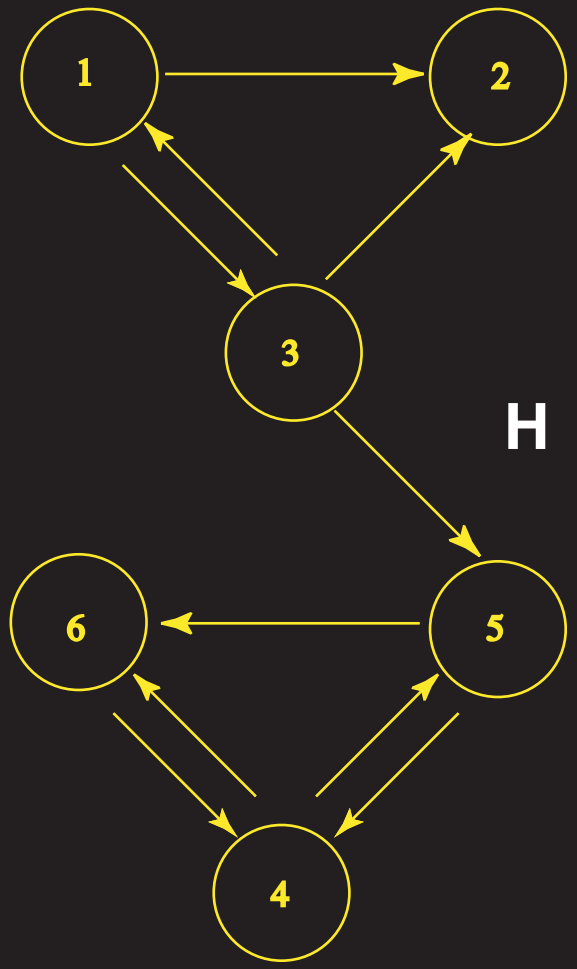
$$H = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



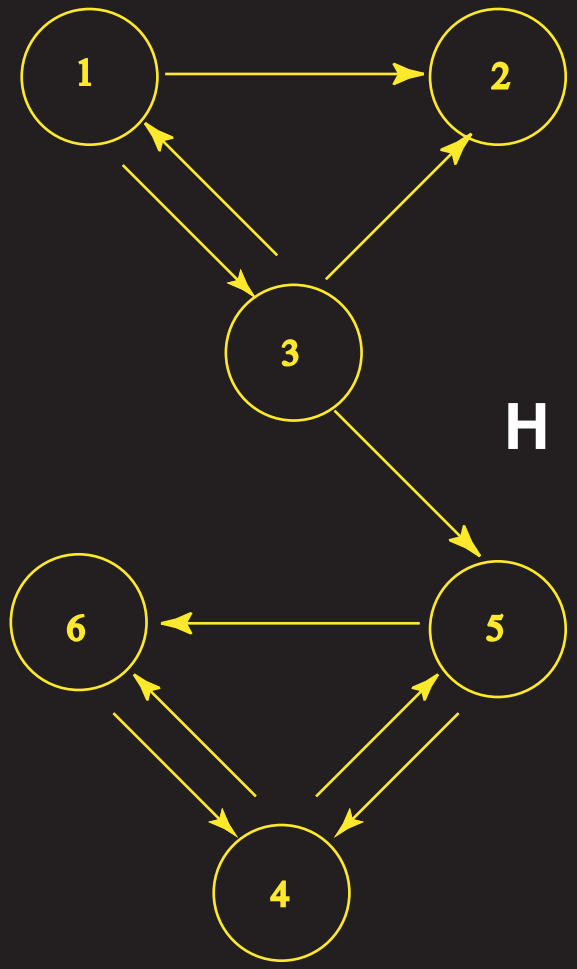
$$H = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



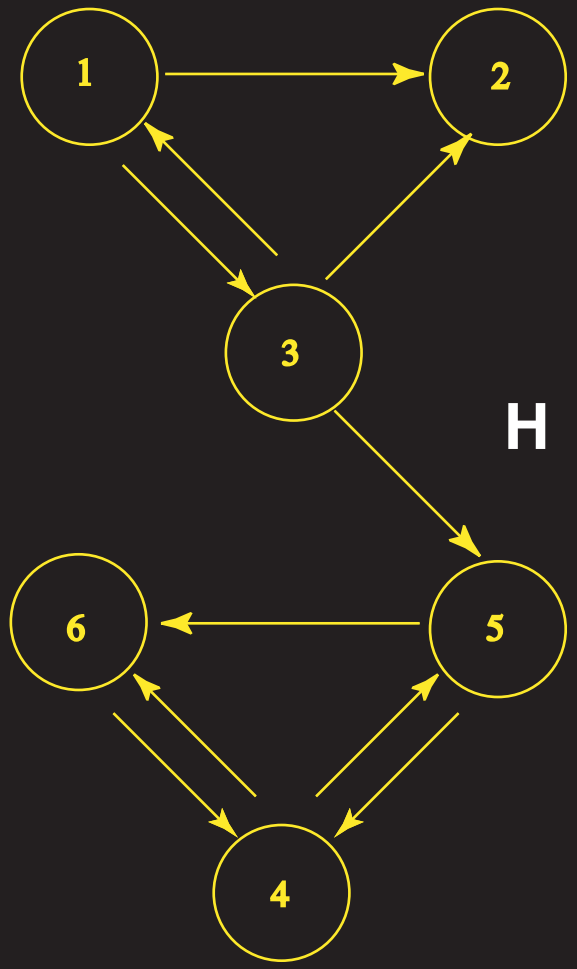
$$H = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ & & & & & \\ & & & & & \\ & & & & &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

Tiny Web



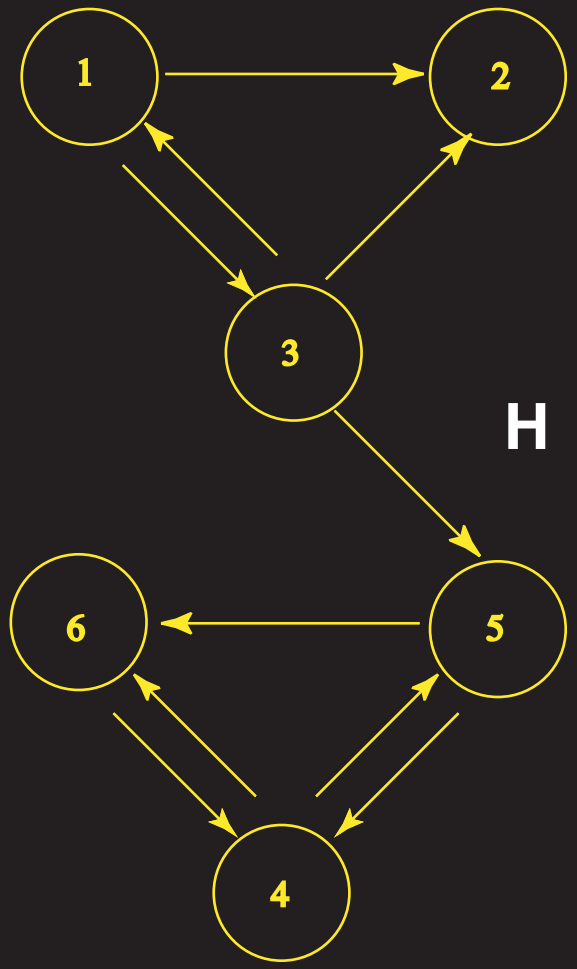
$$H = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ & & &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

Tiny Web



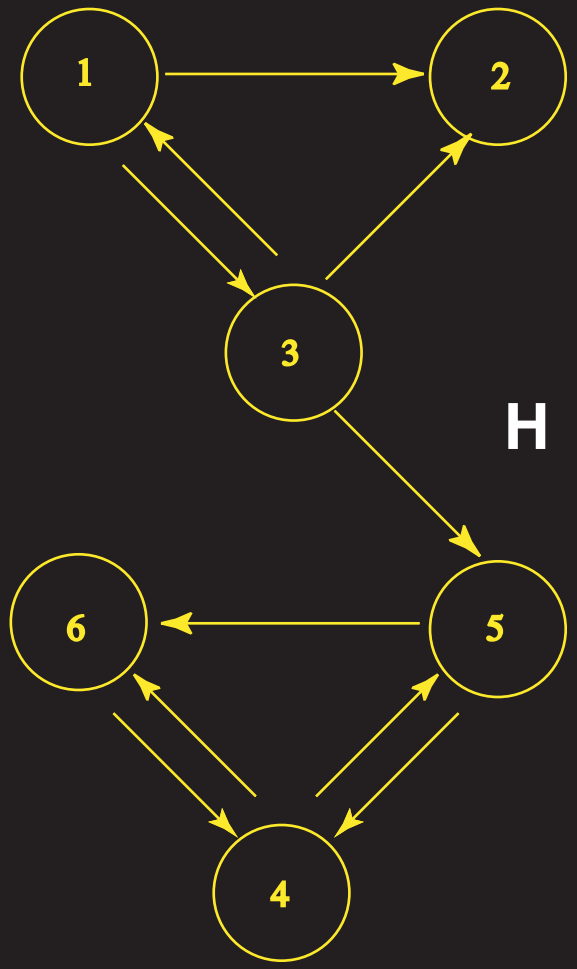
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2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

Tiny Web



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

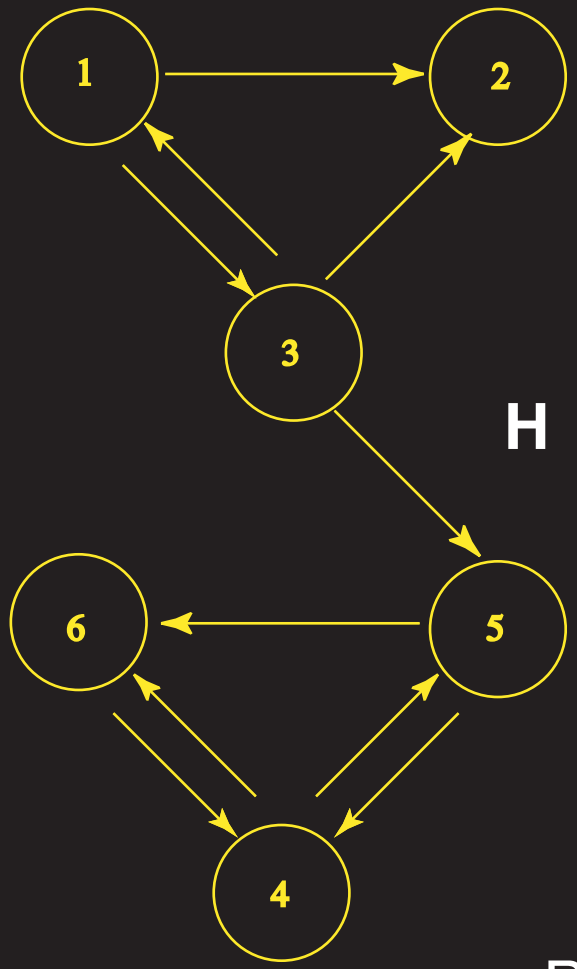
Tiny Web



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

- A random walk on the Web Graph

Tiny Web



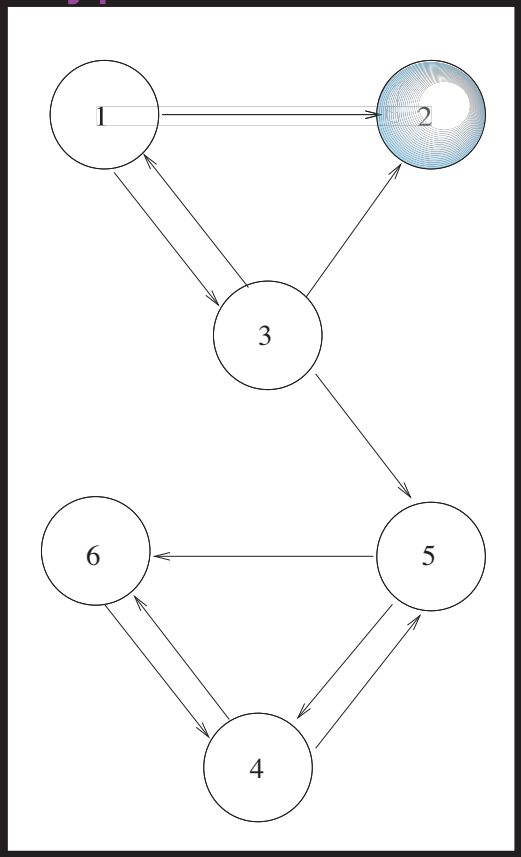
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2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

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

Ranking with a Random Surfer

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

Hyperlink as vote

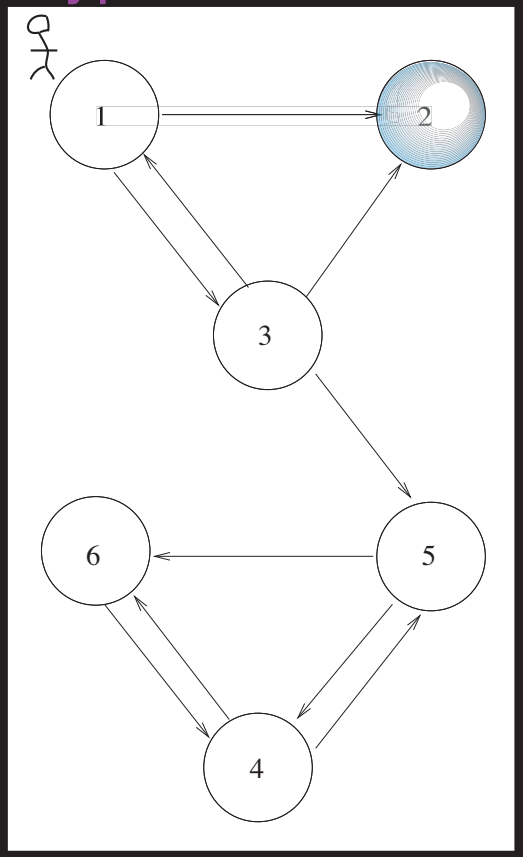


Markov chain

Ranking with a Random Surfer

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

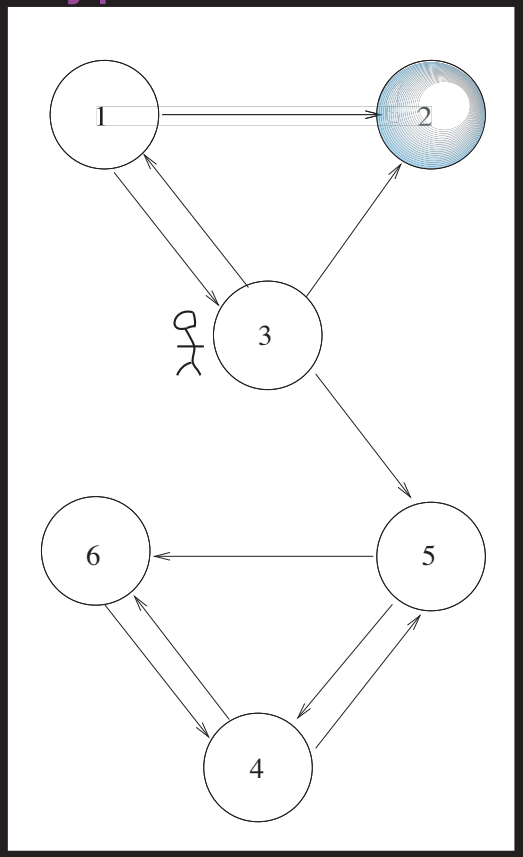
Hyperlink as vote



Ranking with a Random Surfer

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

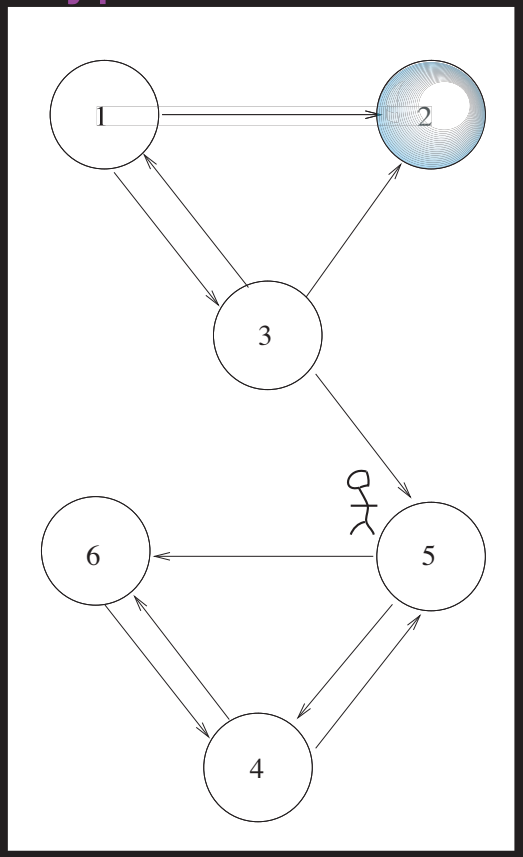
Hyperlink as vote



Ranking with a Random Surfer

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

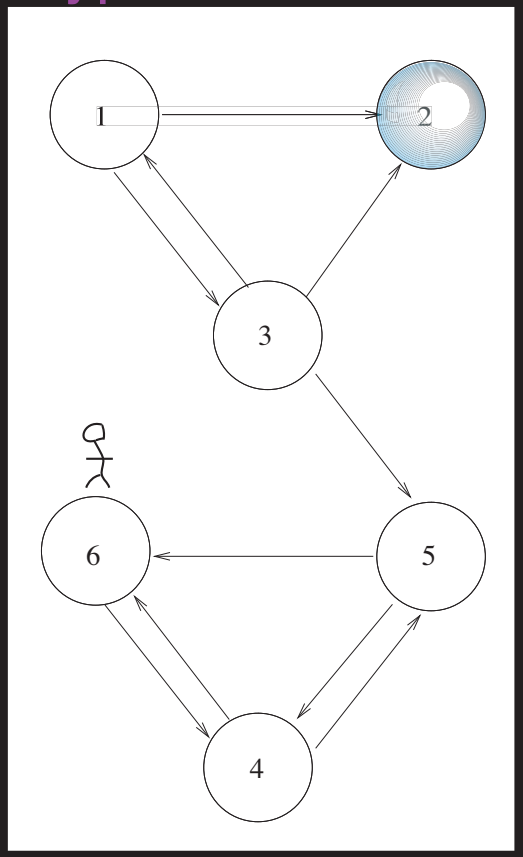
Hyperlink as vote



Ranking with a Random Surfer

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

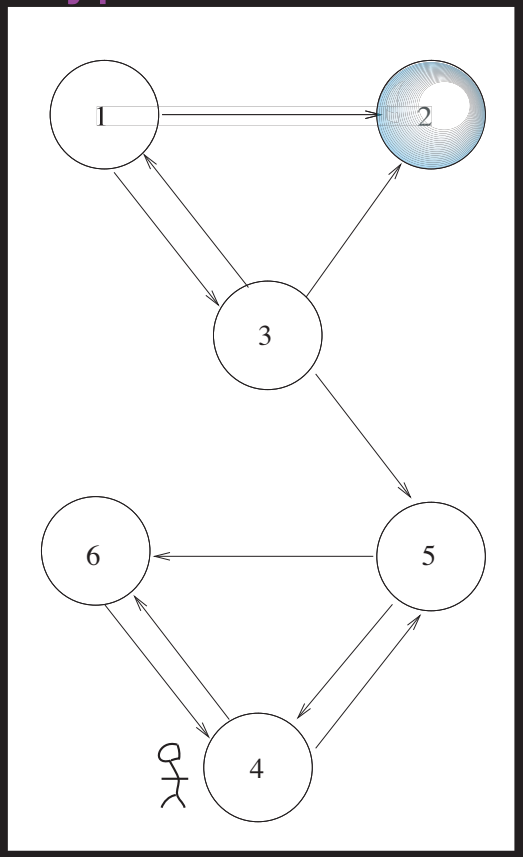
Hyperlink as vote



Ranking with a Random Surfer

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

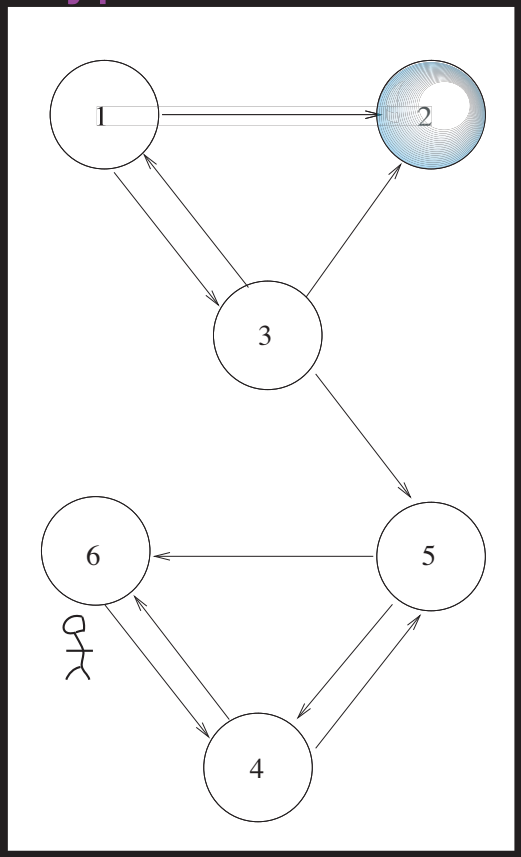
Hyperlink as vote



Ranking with a Random Surfer

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

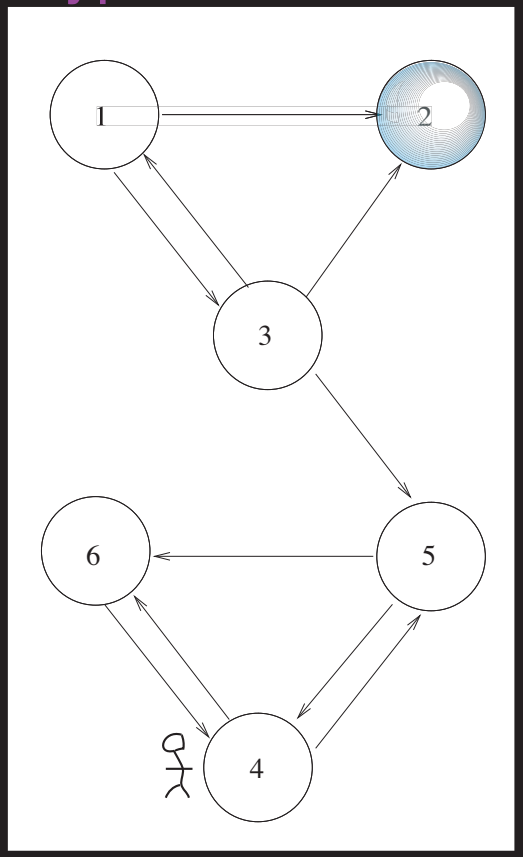
Hyperlink as vote



Ranking with a Random Surfer

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

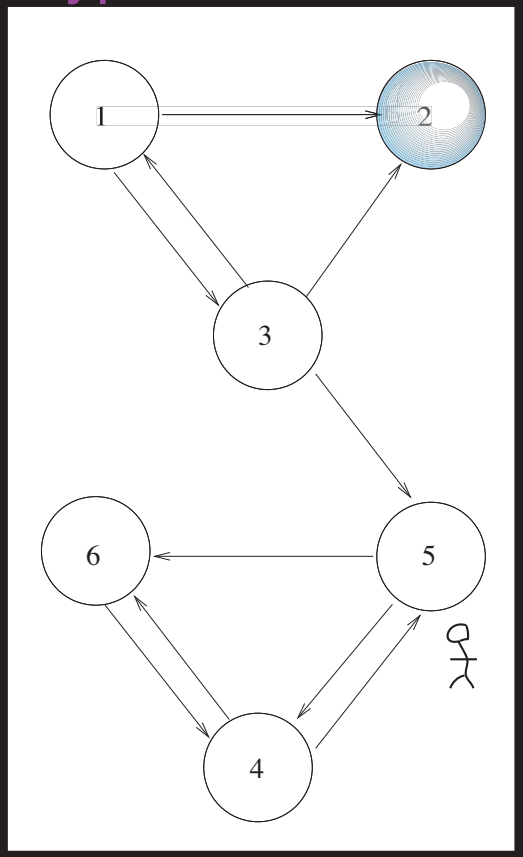
Hyperlink as vote



Ranking with a Random Surfer

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

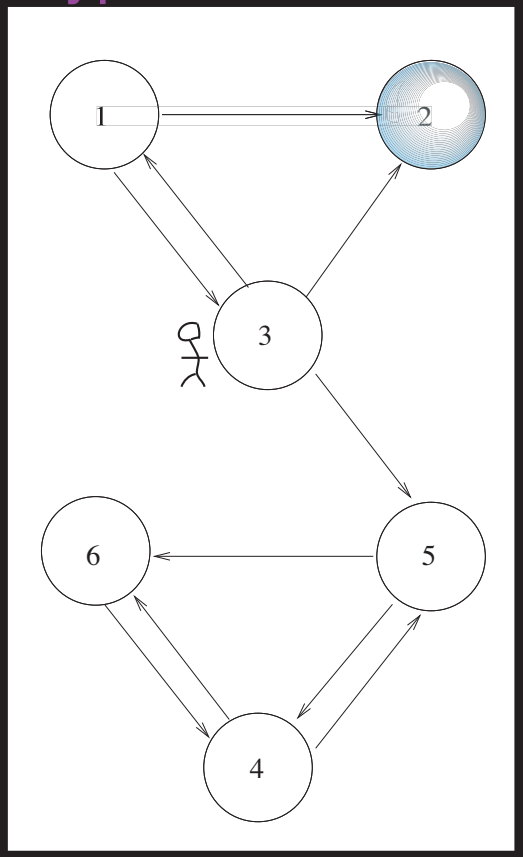
Hyperlink as vote



Ranking with a Random Surfer

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

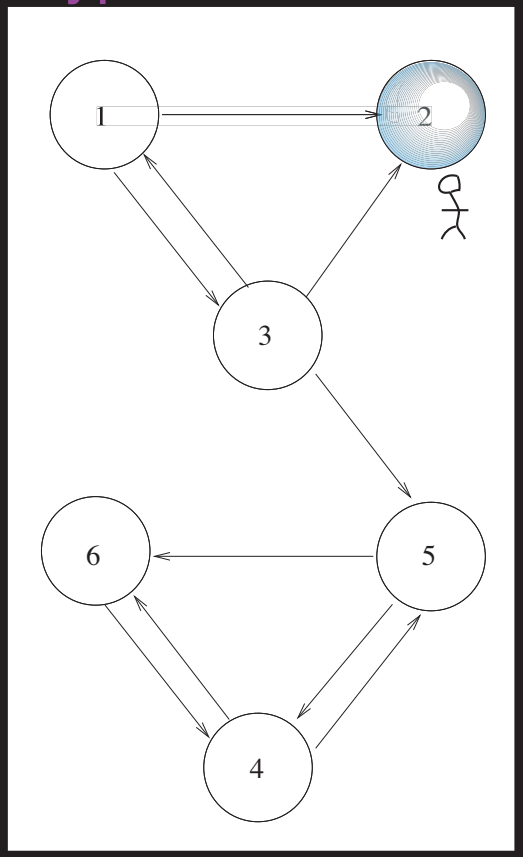
Hyperlink as vote



Ranking with a Random Surfer

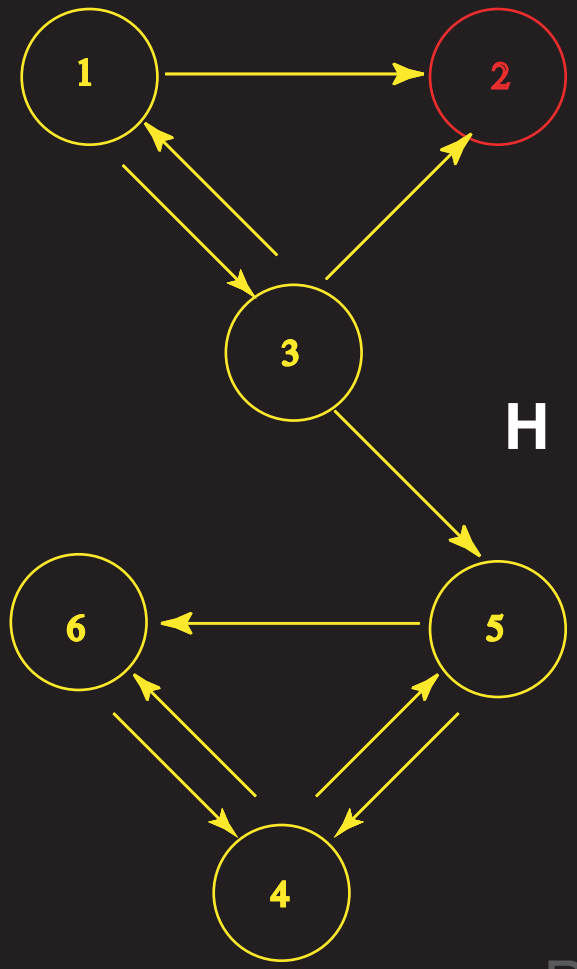
- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote



page 2 is a dangling node

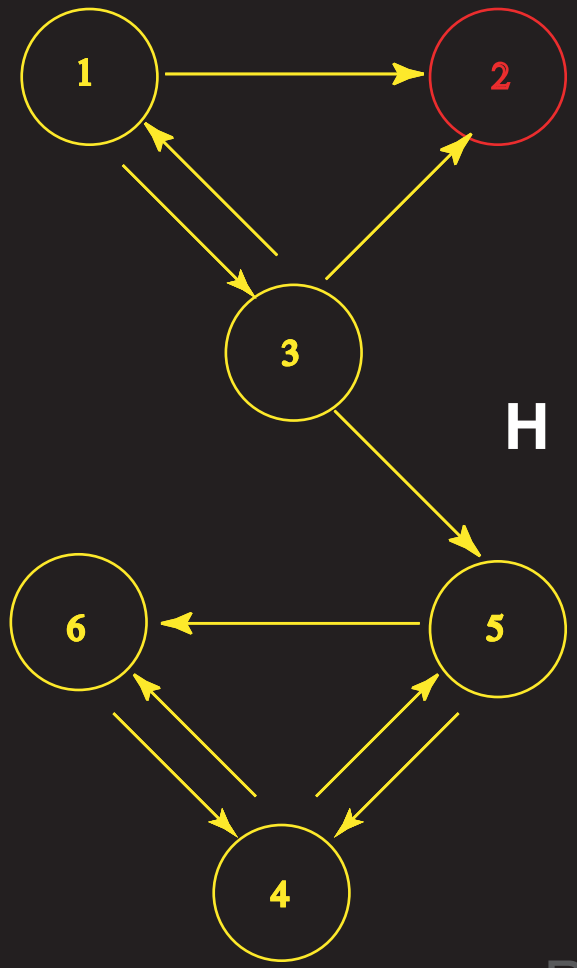
Tiny Web



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

- 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"

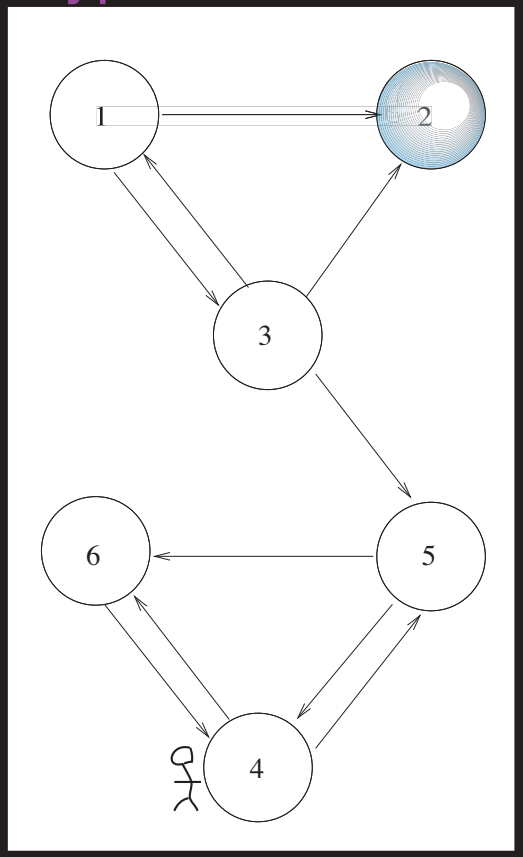
The Fix

Allow Web Surfers To Make Random Jumps

Ranking with a Random Surfer

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

Hyperlink as vote



surfer “teleports”

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 Is Not Strongly Connected

Nasty Problem

The Web Is Not Strongly Connected

- S is reducible

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

Nasty Problem

The Web Is Not Strongly Connected

∴ S is reducible

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

- 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 $(P_i \to P_j \to P_i)$

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 $(P_i \rightarrow P_j \rightarrow P_i)$

- 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 $(P_i \to P_j \to P_i)$

- 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 $\mathbf{G}$

Ranking with a Random Surfer

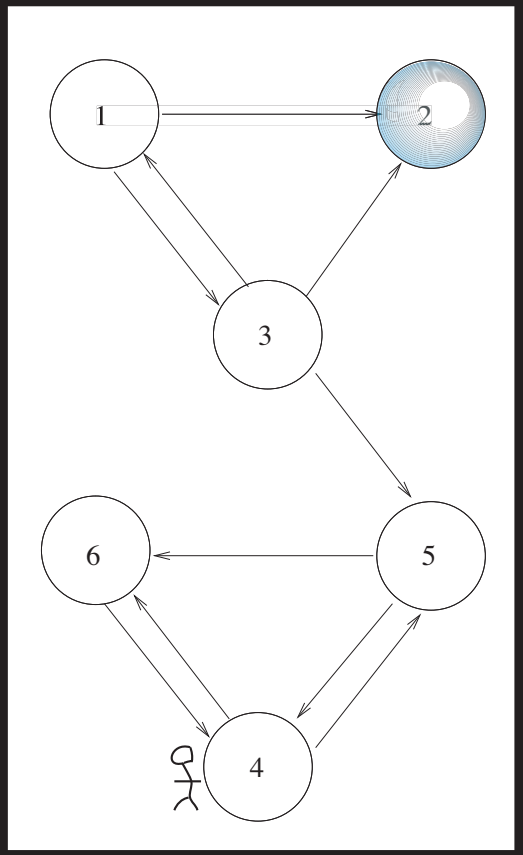
- If a page is “important,” it gets lots of votes from other important pages, which means the random surfer visits it often.
- Simply count the number of times, or $proportion\ of\ time$, the surfer spends on each page to create ranking of webpages.

Ranking with a Random Surfer

- If a page is “important,” it gets lots of votes from other important pages, which means the random surfer visits it often.
- Simply count the number of times, or *proportion of time*, the surfer spends on each page to create ranking of webpages.

Proportion of Time

Page 1 = .04
Page 2 = .05
Page 3 = .04
Page 4 = .38
Page 5 = .20
Page 6 = .29



Ranked List of Pages

Page 4
Page 6
Page 5
Page 2
Page 1
Page 3

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

Search Issues

Spamming

- Link Farms

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)

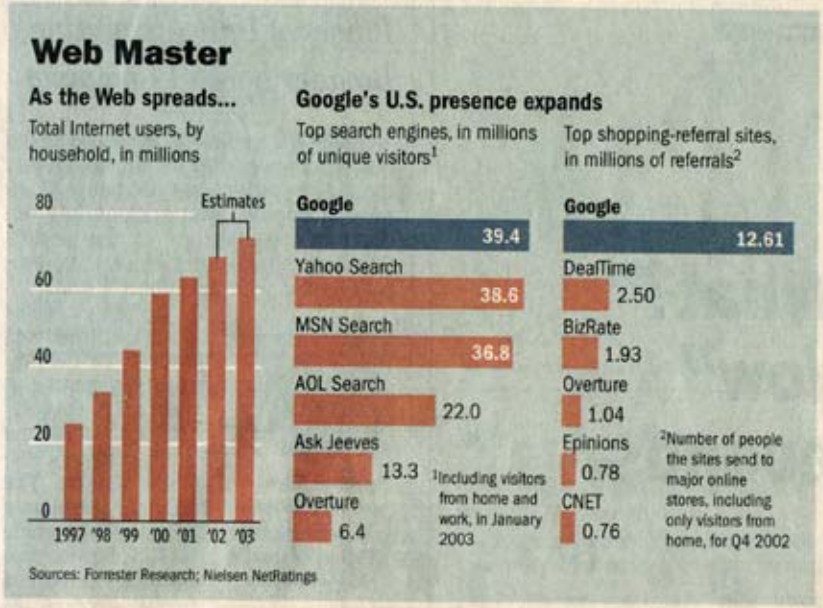
Turkey's parliament debated legislation to let the U.S. deploy 62,000 to open a northern front. Kurdish soldiers lined roads in a show of force as U.S. officials traveled into Iraq's north for an opposition conference.

* * *

■ **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 Goo-

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.

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

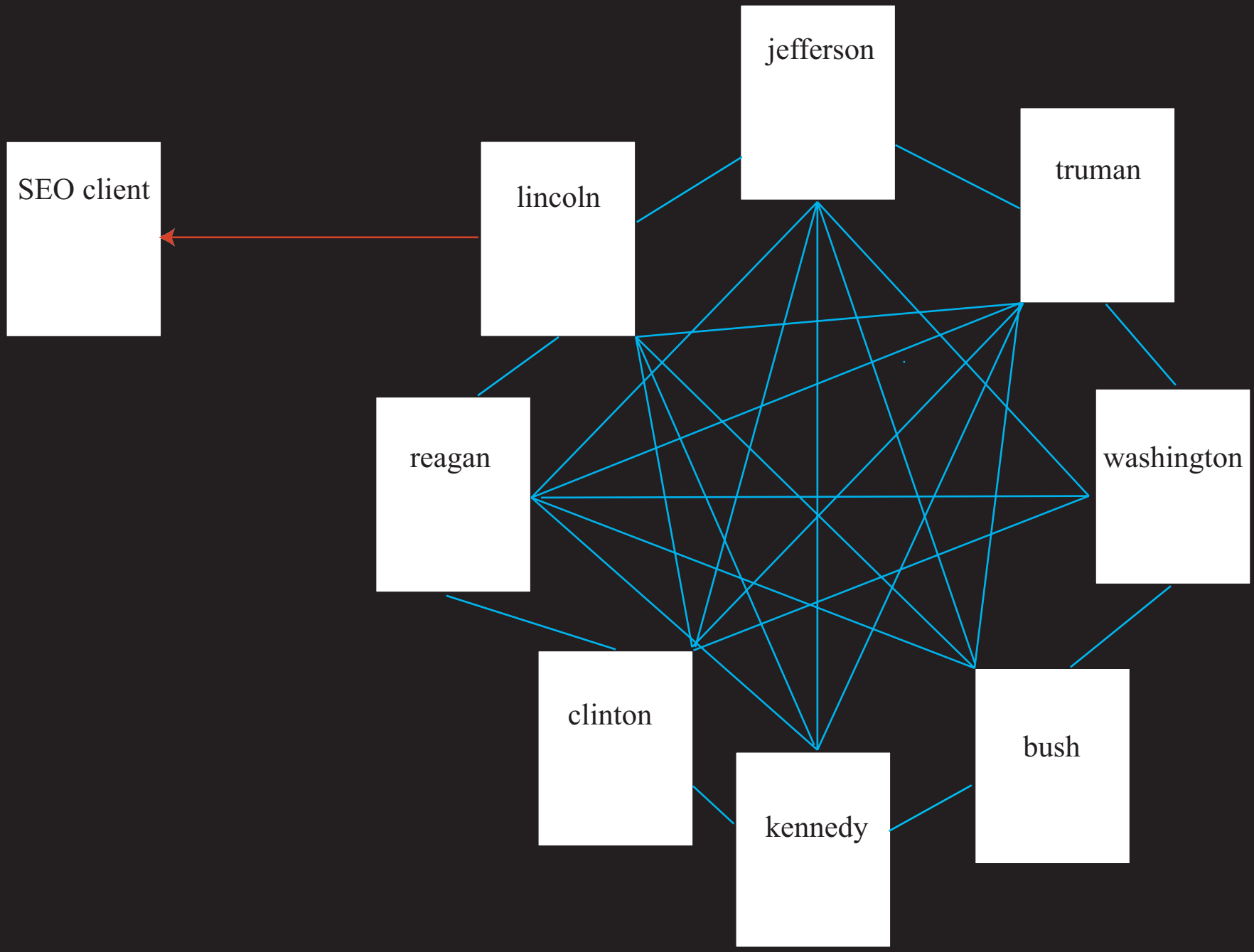
Fiat Patria Is Set to Bec

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 uni

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

Link Farms


jefferson
SEO client
lincoln
truman
reagan
washington
clinton
bush
kennedy

Search Issues

Spamming

- Link Farms
- Google Bombs

News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
Business
Health
Science/Nature
Technology
Entertainment
Have Your Say
Country Profiles
In Depth
Programmes

RELATED SITES
BBC SPORT
BBC WEATHER
BBC ON THIS DAY

LANGUAGES
ESPAÑOL
BRASIL
CARIBBEAN

Last Updated: Sunday, 7 December, 2003, 15:04 GMT

E-mail this to a friend Printable version

'Miserable failure' links to Bush

George W Bush has been Google bombed.


Bush has been the target of similar pranks before

Web users entering the words "miserable failure" into the popular search engine are directed to the biography of the president on the White House website.

The trick is possible because Google searches more than just the contents of web pages - it also counts how often a site is linked to, and with what words.

Thus, members of an online community can affect the results of Google searches - called "Google bombing" - by linking their sites to a chosen one.

Weblogger Adam Mathes is credited with inventing the practice in 2001, when he used it to link the phrase "talentless hack" to a friend's website.

The search engine can be manipulated by a fairly small group of users, one report suggested.

Newsday newspaper says as few as 32 web pages with the words "miserable failure" link to the Bush biography.

The Bush administration has been on the receiving end of pointed Google bombs before.

> **"If you are George Bush and typed the country's name in the address bar, make sure that it is spelled correctly (IRAQ)"**
>
> Prank website

In the run-up to the Iraq war, internet users manipulated Google so the phrase "weapons of mass destruction" led to a joke page saying "These Weapons of Mass Destruction cannot be displayed."

The site suggests "clicking the regime change button", or "If you are George Bush and typed the country's name in the address bar, make sure that it is spelled correctly (IRAQ)".

E-mail this to a friend Printable version

LINKS TO MORE AMERICAS STORIES

Select Go

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

RELATED INTERNET LINKS:
White House
Google bombing
The BBC is not responsible for the content of external internet sites

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

E-mail services | **Desktop ticker** | **Mobiles/PDAs** |

Back to top ^^

News Front Page | **Africa** | **Americas** | **Asia-Pacific** | **Europe** | **Middle East** | **South Asia**
UK | **Business** | **Entertainment** | **Science/Nature** | **Technology** | **Health**
Have Your Say | Country Profiles | In Depth | Programmes

BBCi Homepage >> | BBC Sport >> | BBC Weather >> | BBC World Service >>

ABOUT BBC NEWS | Help | Feedback | News sources | Privacy | About the BBC

BLAH3.COM BLAHBLAHBLAHBLAHBLAHBLAH
CH FOR THE NEW CENTURY
Dusty & Yellowing - The Blah3 Archives
Complaints, compliments, arguments? Email me

'Den of Thieves'
The ? Campaign, 2002
'Fair & Balanced' Day

Blahroll

A Level Gaze
A Skeptical Blog
Ain't No Bad Dude
Angry Bear
Ann Slanders
Apathy, Inc.
Army of Fun
Atrios
Attorney At Arms
Avedon's Sideshow
Bag Times
BartCop E!
BartCop!
Bellum Americanum
Big Picnic
Bitter Obscurity
Booknotes
Bunsen
Burgblog
Bush Is A Moron
BushFlash
BushLiar
BusyBusyBusy
Byrd's Brain
Certain Shade of Green
Chimes at Midnight
Chris Nelson
Circumspect
CNN Lies
Conniption
Counterspin
Cursor
Daily Brew
Daily Cynic
Daily Kos
Daily Outrage
Daily War News
Damfacrats
Deckie Holmes
Democratic Veteran
Dodona
dratfink
Duckwing
E Pluribus Unum
Estimated Prophet
Ethel
Federal Examiner
Fengi
For Freedom Century
Frog'N'Blog
Ge. JC Christian
GeekPol
Genoan Sailor
GeoDog
Get Donkey!

[<< "Happy Ramadan, y'all..."] [Main Index] [>>"His heart just isn't in it...."]

10/27/2003 Archived Entry: "I'm taking part in a new web project..."

I'm taking part in a new web project...

From this day forth, I will refer to George W. Bush as a Miserable Failure at least once a day. Why, you ask? Well, someone came up with this great idea to link George W. Bush and Miserable Failure in popular search engines. If you have a blog or web site, help raise the link between George W. Bush and the phrase 'miserable failure' by copying this link and placing somewhere on your site or blog.

Thank you very much for your participation.

Replies: 16 people speak up

Great idea!

Posted by rlrr @ 10/27/2003 10:06 PM NY

That is genius. I could add a few other keywords, like "pathetic". I will post it on my blog now...

Posted by Political Pulpit @ 10/28/2003 02:32 PM NY

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

Stay tuned...

Posted by Drewcifer @ 10/28/2003 02:35 PM NY

Done!

Posted by Maru @ 10/28/2003 08:46 PM NY

thats great, another thing I think
might be good to use: tax cuts for the wealthy....welfare for the wealthy. just my 2 cents.

Posted by doodaa @ 10/29/2003 03:01 AM NY

Call me a liberal lemming, I guess. ;) I'm in.

Posted by BJ @ 10/29/2003 09:28 AM NY

The key is stating it in connection with terms that will be widely searched. It does no good to simply say "George Bush is a miserable failure" because no one will ever search for that. It might be fun at a parties to show how often the two are in the same sentence in a Google search, but otherwise it does little to advance the theme.

What will work is connecting it to frequent search times, such as "Iraq policy". For instance "George Bush's Iraq Policy is a miserable failure."

The plan shouldn't be to link Miserable Failure to George Bush, but to link Miserable Failure to George Bush and two or three choice, frequently searched phrases.

Overture.com has a keyword suggestion tool that shows how many times certain terms are coming up in searches. Using that tool, I can determine that in September the search for "bush george iraq saddam" gets about 12 times more queries than "george bush iraq speech". "george bush biography" gets a huge amounts of hits compared to something like "george bush policy".

So someone needs to write about three complete sentences using these terms based on verfiable search results and including the "miserable failure" phrase and then advocate for that exact usage.

According to Overture, the phrases "george Bush miserable failure" were not queried even once in their sample during the month just passed.

Posted by Joe Briefcase @ 10/29/2003 10:51 AM NY

how about drunken, illiterate, mendacious, runt-like miserable failure?

Posted by tim @ 10/29/2003 11:58 AM NY

Hahaha, that's very productive. This is why everyone knows that liberals are stupid. They do stupid things.

Posted by Reek Stankleberry @ 10/29/2003 12:04 PM NY

how about, instead of calling it lies--anyone can lie--how about calling it HORSEFEATHERS AND CODSWALLOP! Pin that on him too.

Google Advanced Search Preferences Language Tools Search Tips

miserable failure

Google Search

Web Images Groups Directory News

Searched the web for **miserable failure**. Results **1** - **10** of about **257,000**. Search took **0.08** seconds.
Tip: In most browsers you can just hit the return key instead of clicking on the search button.

Michael Moore.com
Wednesday, January 14th, 2004 I'll Be Voting For Wesley Clark /
Good-Bye Mr. Bush — by Michael Moore. Many of you have written **...**
Description: Official site of the gadfly of corporations, creator of the film Roger and Me and the television show...
Category: Arts > Celebrities > M > Moore, Michael
www.michaelmoore.com/ - 43k - Cached - Similar pages

Biography of President George W. Bush
Home > President > Biography President George W. Bush En Español.
George W. Bush is the 43rd President of the United States. He **...**
Description: Biography of the president from the official White House web site.
Category: Kids and Teens > School Time > ... > Bush, George Walker
www.whitehouse.gov/president/gwbbio.html - 29k - Cached - Similar pages

Biography of Jimmy Carter
Home > History & Tours > Past Presidents > Jimmy Carter. Jimmy Carter.
Jimmy Carter aspired to make Government "competent and compassionate **...**
Description: Short biography from the official White House site.
Category: Society > History > ... > Presidents > Carter, James Earl
www.whitehouse.gov/history/presidents/jc39.html - 36k - Cached - Similar pages

Senator Hillary Rodham Clinton: Online Office Welcome Page
Dear Friend,. Thank you for visiting my on-line office! I appreciate
your interest in the issues before the United States Senate. **...**
Description: Official US Senate web site of Senator Hillary Rodham Clinton (D - NY).
Category: Society > History > ... > First Ladies > Clinton, Hillary
clinton.senate.gov/ - 9k - Cached - Similar pages

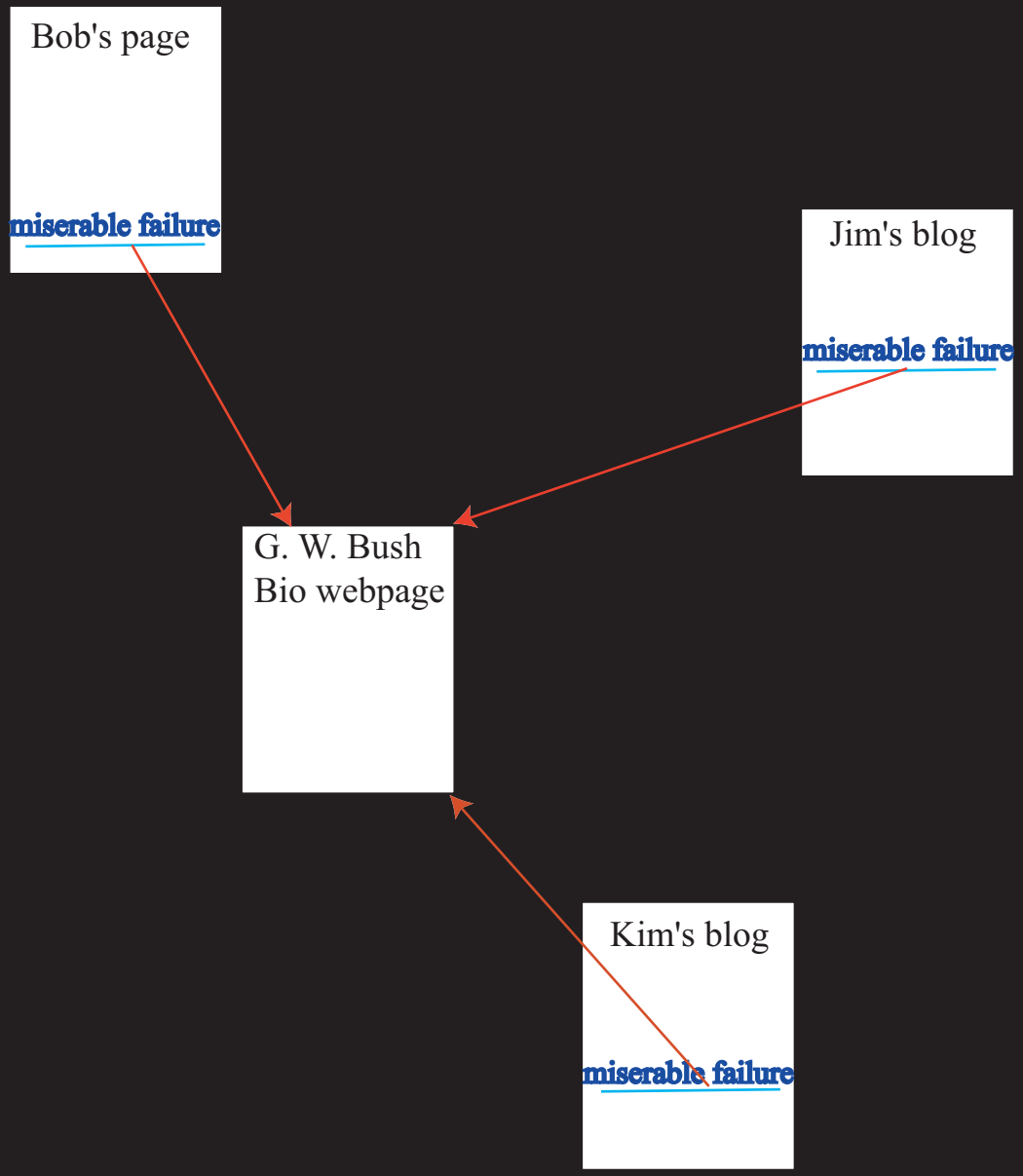
BBC NEWS | Americas | '**Miserable failure**' links to Bush
'**Miserable failure**' links to Bush. **...** Prank website. Newsday newspaper says as few as 32 web pages with the words "**miserable failure**" link to the Bush biography. **...**
news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - Cached - Similar pages

Atlantic Unbound | Politics & Prose | 2003.09.24
... Atlantic Unbound | September 24, 2003 Politics & Prose | by Jack Beatty
"A **Miserable Failure**" Will Bush be re-elected? Only if voters **...**
www.theatlantic.com/unbound/polipro/pp2003-09-24.htm - 22k - Cached - Similar pages

miserable failure | Hillary Clinton | Hildebeest
... Miserable Failure. Quotes for the History Books. **...** You may also want to check out the **Miserable Failure** Project. and the cuckolded dyke Project. and the **...**
miserable-failure.blogspot.com/ - 60k - Cached - Similar pages

Dick Gephardt for President - Welcome
... to preserve some large part of the Bush tax cut. I think retaining

Google Bomb


Bob's page
miserable failure
Jim's blog
miserable failure
G. W. Bush
Bio webpage
Kim's blog
miserable failure

Search Issues

Spamming

- Link Farms
- Google Bombs

Personalization

- Google's psearch, A9, Kartoo

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.

Jason Raish

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


KartOO
www.kartoo.com
help
<<
langville
Search
english pages
options
Products
Saved map
New map
Topics :
book google
updating pagerank
engine rankings
carl meyer
langville and carl
langville and meyer
survey of eigenvector metho
list of publications
markov chains
assistant professor
mathematics department
meyer
google
beyond
description
science
pagerank
book
publications
carl
abstract
information
retrieval
www.cofc.edu
members.lycos.co.uk
www.krellinst.org
abstract
information
retrieval
hard
publications
www.informatik.uni-trier.de
meyer.math.ncsu.edu
pagerank
carl
www.internetmathematics.org
www.allbookstores.com
book
meyer
press.princeton.edu
www.pupress.princeton.edu
google
beyond
description
science
www.nni-news.com
www.langreiter.com
www.realestate.com.au
print the map
Send a map
Add a site
Add a Topic
save the map...
Preview by
THUMBSHOTS
KartOO
TECHNOLOGIES
next map
18 700 Found results
25 - 40

Search Issues

Spamming

- Link Farms
- Google Bombs

Personalization

- Google's psearch, A9, Kartoo

Privacy

- AOL Data Leak

A Face Is Exposed for AOL Searcher No. 4417749

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

SIGN IN TO E-MAIL THIS
PRINT
SINGLE PAGE
REPRINTS
SAVE

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

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Multimedia

Graphic: What Revealing Search Data Reveals

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

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

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

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

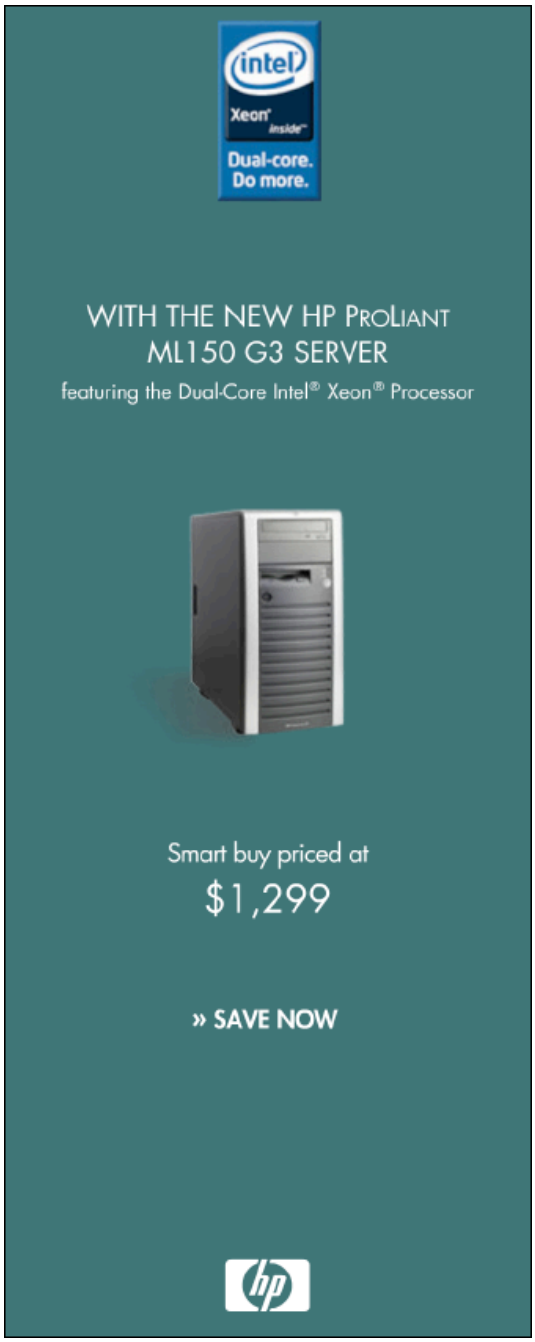
But the detailed records of searches conducted by Ms. Arnold and 657,000 other Americans, copies of which continue to circulate online, underscore how much people unintentionally reveal about themselves when they use search engines — and how risky it can be for companies like AOL, Google and Yahoo to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

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

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

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

More Articles in Technology »



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

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

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

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

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

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

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

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

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

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

1 | 2 NEXT PAGE »

Saul Hansell contributed reporting for this article.

More Articles in Technology »

Related Articles

How to Digitally Hide (Somewhat) in Plain Sight (August 12, 2006)
Your Life as an Open Book (August 12, 2006)
AOL Removes Search Data On Vast Group Of Web Users (August 8, 2006)
U.S. Wants Internet Companies to Keep Web-Surfing Records (June 2, 2006)

Related Searches

America Online Inc
Privacy
Computers and the Internet
Google Inc

INSIDE NYTIMES.COM

TimesSelect

Vinocur: A Price for Denying Holocaust?

N.Y. / REGION »

In Douglass Tribute, Folklore and Fact Collide

U.S. »

A Dose of Maturity for a California Protest

TimesSelect

Bush's Gold-Plated Indifference

Additional notes and reader comments on Paul Krugman's column on health care.

MOVIES »

Documenting an Obsession and a Marriage

BOOKS »

Robert Loomis's Career in Letters

Search Issues

Spamming

- Link Farms
- Google Bombs

Personalization

- Google's psearch, A9, Kartoo

Privacy

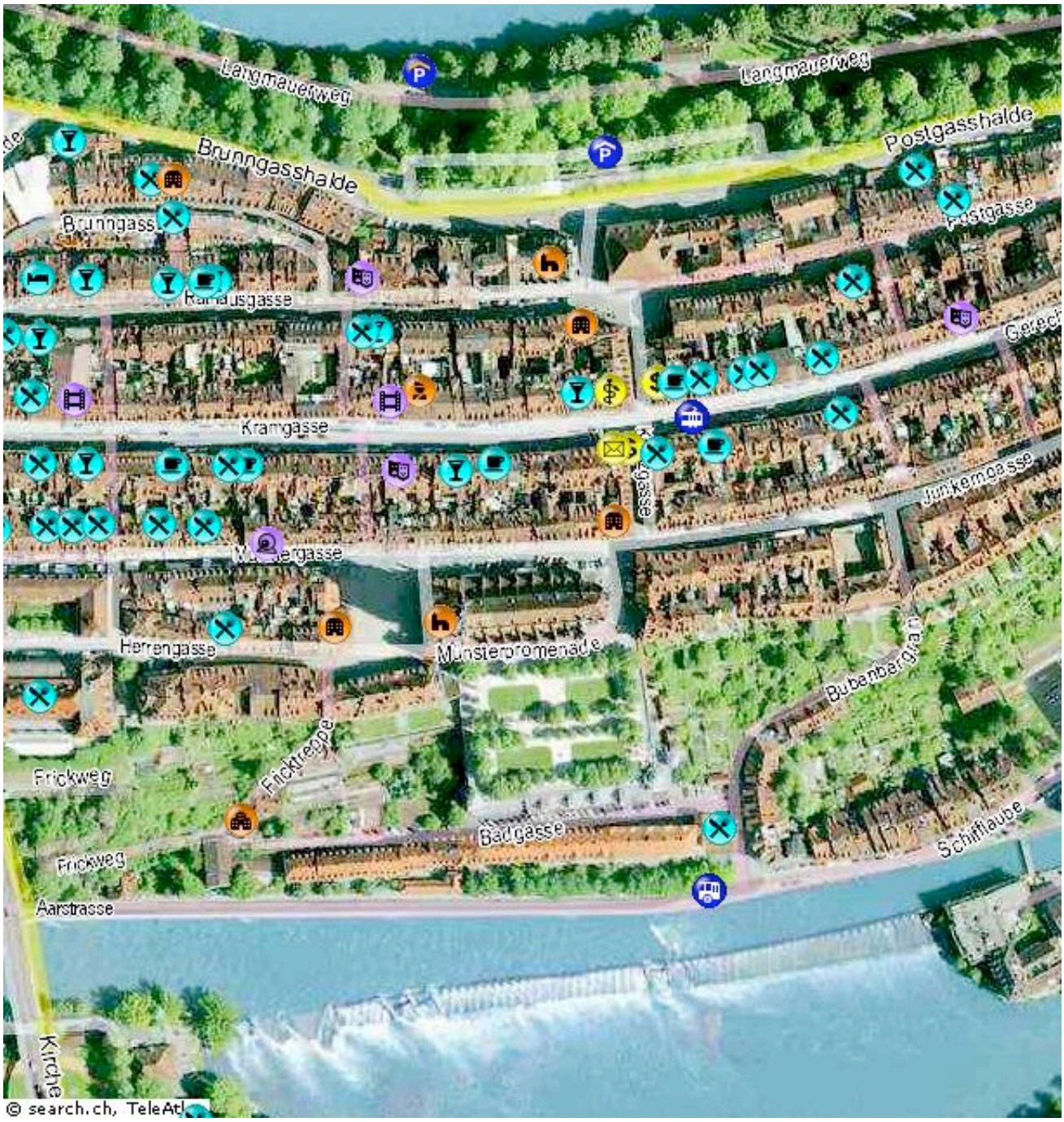
- AOL Data Leak

Data Fusion

- Search.ch

[map.search.ch]

Map: Bern



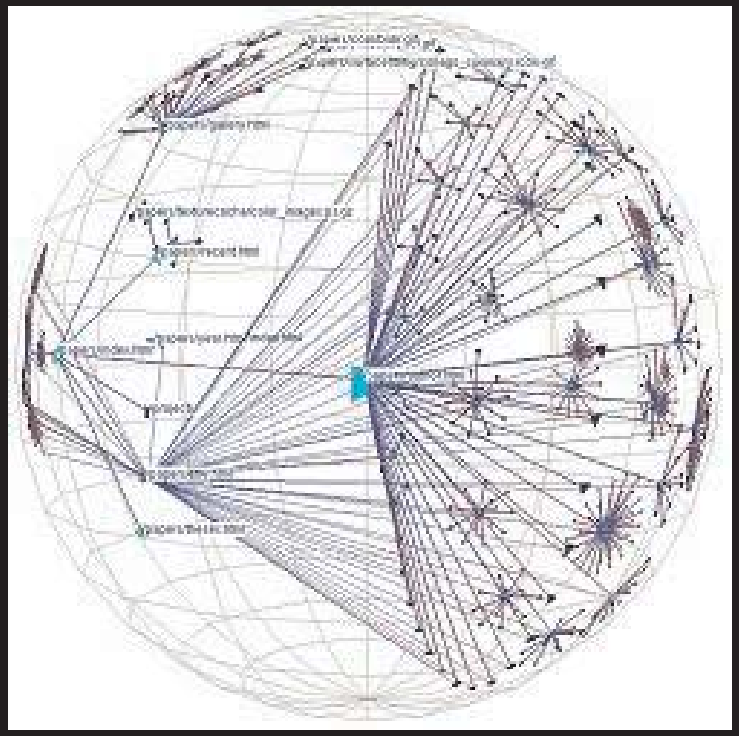
Conclusion

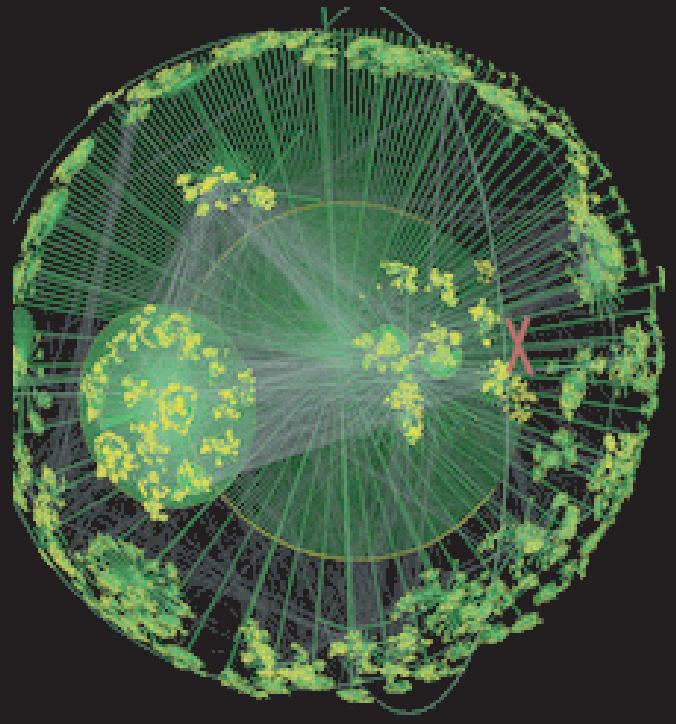
- Link-based scores (PageRank, HITS, etc.) are combined with content scores for final rankings
- Link analysis has dramatically improved search.
- Many continuing CSC and MATH challenges.

Web Graphs

CSC and MATH challenges (problems of scale!)

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





Conclusion

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

Conclusion

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

Elegant and Exciting Application of Linear Algebra

That is Changing the World