Web Search
and Beyond

Carl Meyer

Department of Mathematics
North Carolina State University
Raleigh, NC

Institute Of Advanced Analytics
Raleigh, NC
April 14, 2010

Short History of IR

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

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

System for the Mechanical Analysis and Retrieval of Text

Harvard 1962 – 1965

Cornell 1965 – 1970

Gerard Salton

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

Term–Document Matrices

Start with dictionary of terms

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

Term–Document Matrices

Start with dictionary of terms

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

Index Each Document

Humans scour pages and mark key terms

Term–Document Matrices

Start with dictionary of terms

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

Index Each Document

Humans scour pages and mark key terms

Count f_{ij} = # times term i appears in document j

Term–Document Matrices

Start with dictionary of terms

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

Index Each Document

Humans scour pages and mark key terms

Count f_{ij} = # times term i appears in document j

Term–Document Matrix

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

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?

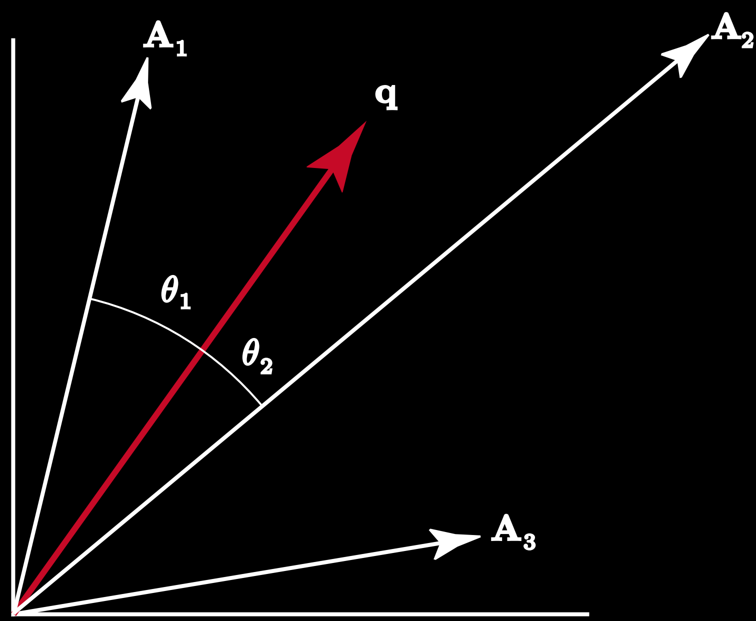
Query Matching

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



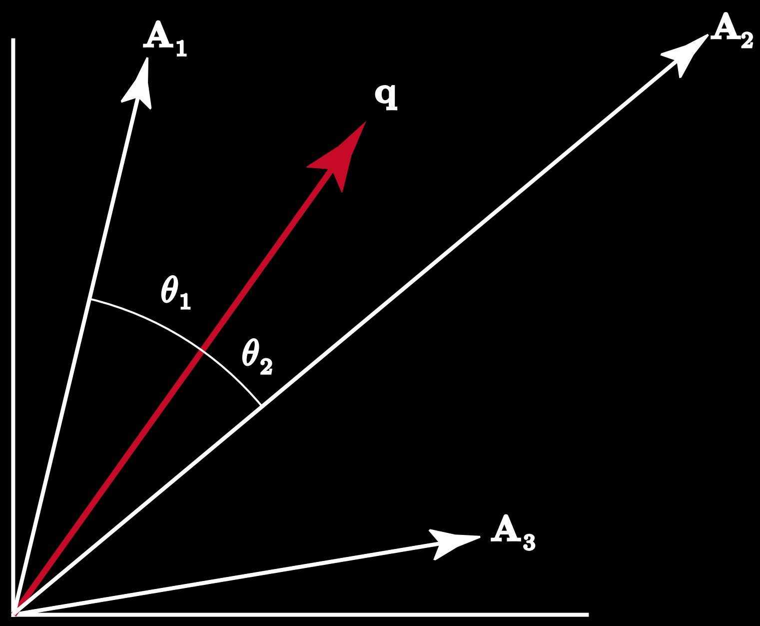
Query Matching

Query Vector

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

How Close is Query to Each Document?

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



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

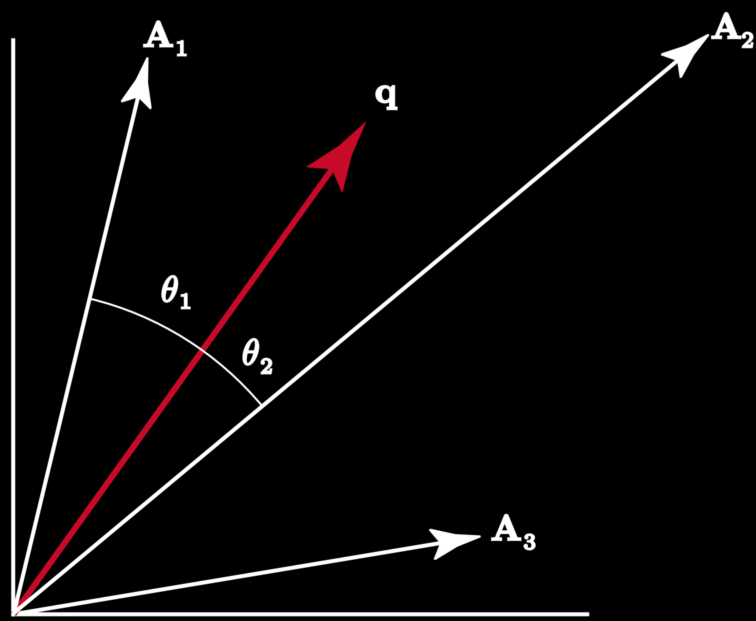
Query Matching

Query Vector

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

How Close is Query to Each Document?

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



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

Rank documents by size of δ_i

Return Document i to user when $\delta_i \geq tol$

Susan Dumais's Improvement

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

- 2 patents for Bell/Telcordia
  - Computer information retrieval using latent semantic structure. U.S. Patent No. 4,839,853, June 13, 1989.
  - Computerized cross-language document retrieval using latent semantic indexing. U.S. Patent No. 5,301,109, April 5, 1994.
- LATENT SEMANTIC INDEXING

Latent Semantic Indexing

Use a finite Fourier expansion of A

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

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

Latent Semantic Indexing

Use a finite Fourier expansion of A

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

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

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

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

Latent Semantic Indexing

Use a finite Fourier expansion of A

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

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

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

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

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

Latent Semantic Indexing

Use a finite Fourier expansion of A

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

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

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

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

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

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

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

Latent Semantic Indexing

Use a finite Fourier expansion of A

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

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

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

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

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

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

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

“Best” mathematical solution

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

Strengths & Weaknesses

Pros

- Finds hidden connections

Strengths & Weaknesses

Pros

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

Strengths & Weaknesses

Pros

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

Strengths & Weaknesses

Pros

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

Cons

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

Strengths & Weaknesses

Pros

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

Cons

- Rankings are query dependent
  - Rank of each doc is recomputed for each query
- Only semantic content used
  - Susceptible to malicious manipulation

Strengths & Weaknesses

Pros

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

Cons

- Rankings are query dependent
  - — Rank of each doc is recomputed for each query
- Only semantic content used
  - — Susceptible to malicious manipulation
- Difficult to add & delete documents

Strengths & Weaknesses

Pros

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

Cons

- Rankings are query dependent
  — Rank of each doc is recomputed for each query
- Only semantic content used
  — Susceptible to malicious manipulation
- Difficult to add & delete documents
- Finding optimal compression requires empirical tuning

Web Documents

Different from other document collections

- It's huge
  - Billions of pages, where average page size $\geq$ 500KB
  - Many-many times the size of Library of Congress print collection

Web Documents

Different from other document collections

- It's huge
  - Billions of pages, where average page size $\geq$ 500KB
  - Many-many times the size of Library of Congress print collection
- It's dynamic
  - 40% of all pages change in a week
  - 23% of .com pages change daily
  - Billions of pages added each year

Web Documents

Different from other document collections

- It's huge
  - Billions of pages, where average page size $\geq$ 500KB
  - Many-many times the size of Library of Congress print collection
- It's dynamic
  - 40% of all pages change in a week
  - 23% of .com pages change daily
  - Billions of pages added each year
- It's self-organized
  - No standards, review process, formats
  - Errors, falsehoods, link rot, and spammers!

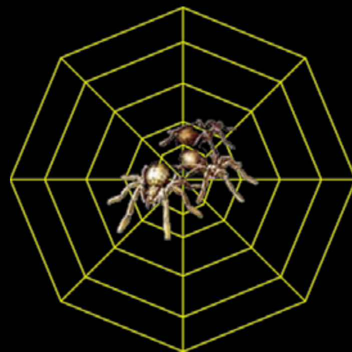
Web Documents

Different from other document collections

- It's huge
  - Billions of pages, where average page size $\geq$ 500KB
  - Many-many times the size of Library of Congress print collection
- It's dynamic
  - 40% of all pages change in a week
  - 23% of .com pages change daily
  - Billions of pages added each year
- It's self-organized
  - No standards, review process, formats
  - Errors, falsehoods, link rot, and spammers!
- It has many users
  - Google alone processes more than 620 million queries per day

Web Search Components

Web Crawlers



Software robots gather web pages

Web Search Components

Web Crawlers

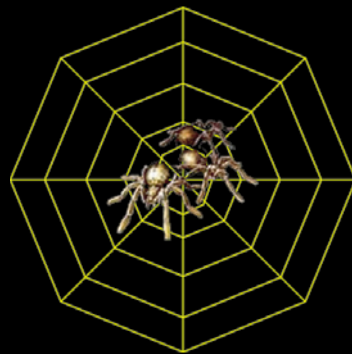
Software robots gather web pages

Doc Server

Stores docs and snippits

Web Search Components

Web Crawlers



Software robots
gather web pages

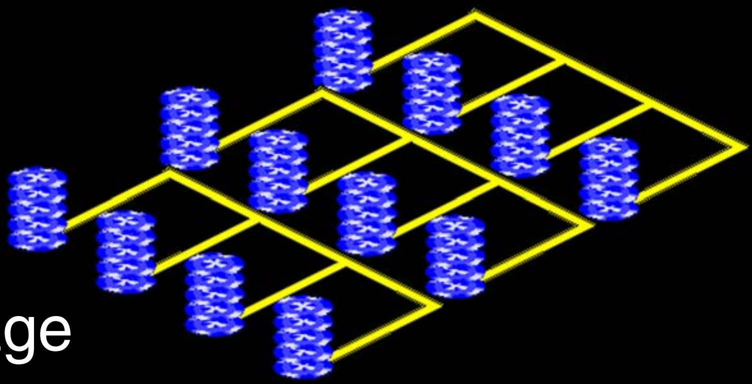
Doc Server

Stores docs
and snippits

Index Server

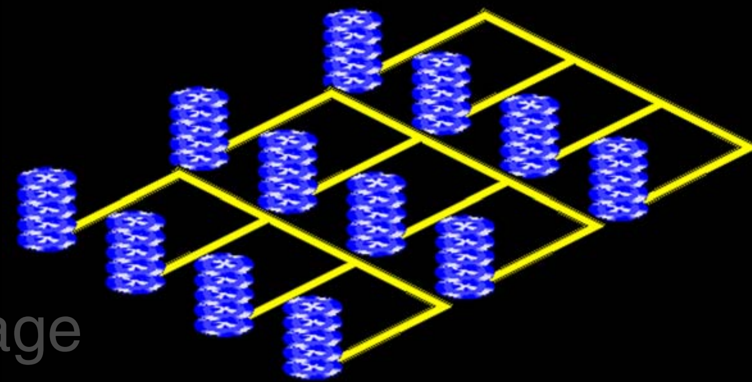
Scans pages and does term indexing
Terms $\longrightarrow$ Pages (similar to book index)

The Ranking Module



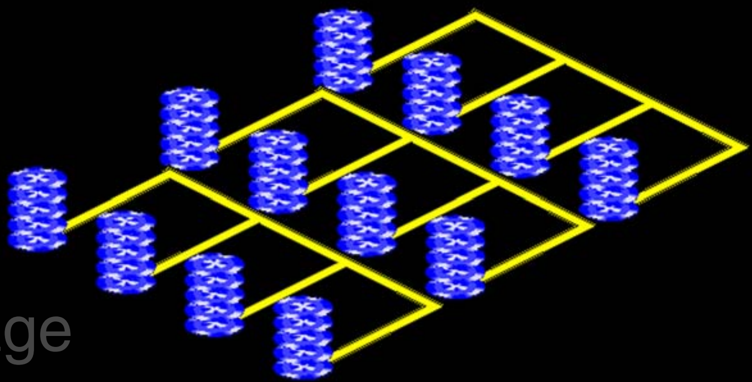
- Measure the importance of each page

The Ranking Module



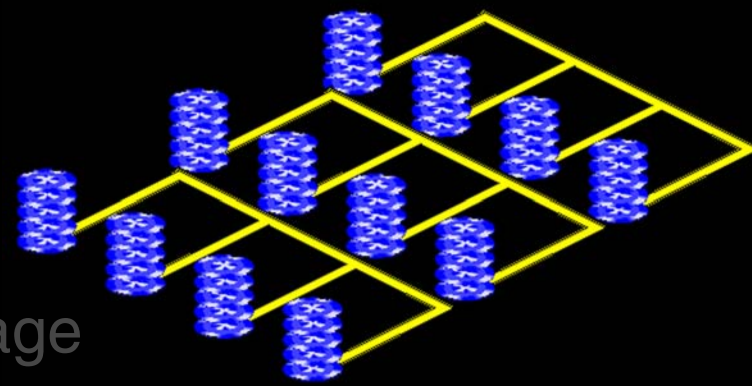
- Measure the importance of each page
- The measure should be Independent of any query

The Ranking Module



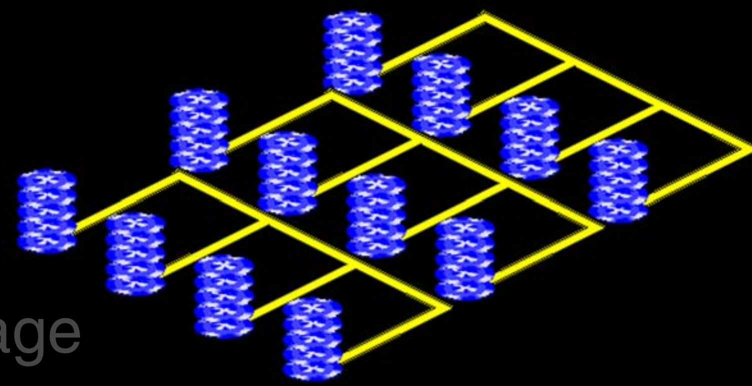
- Measure the importance of each page
- The measure should be Independent of any query
- Compute these measures off-line before queries are processed

The Ranking Module



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

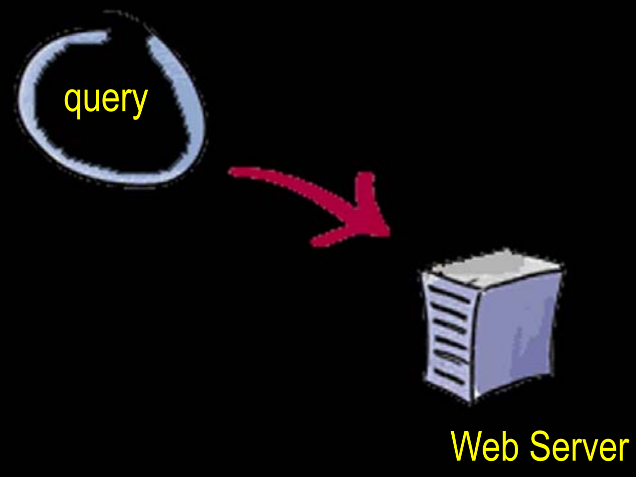
The Ranking Module



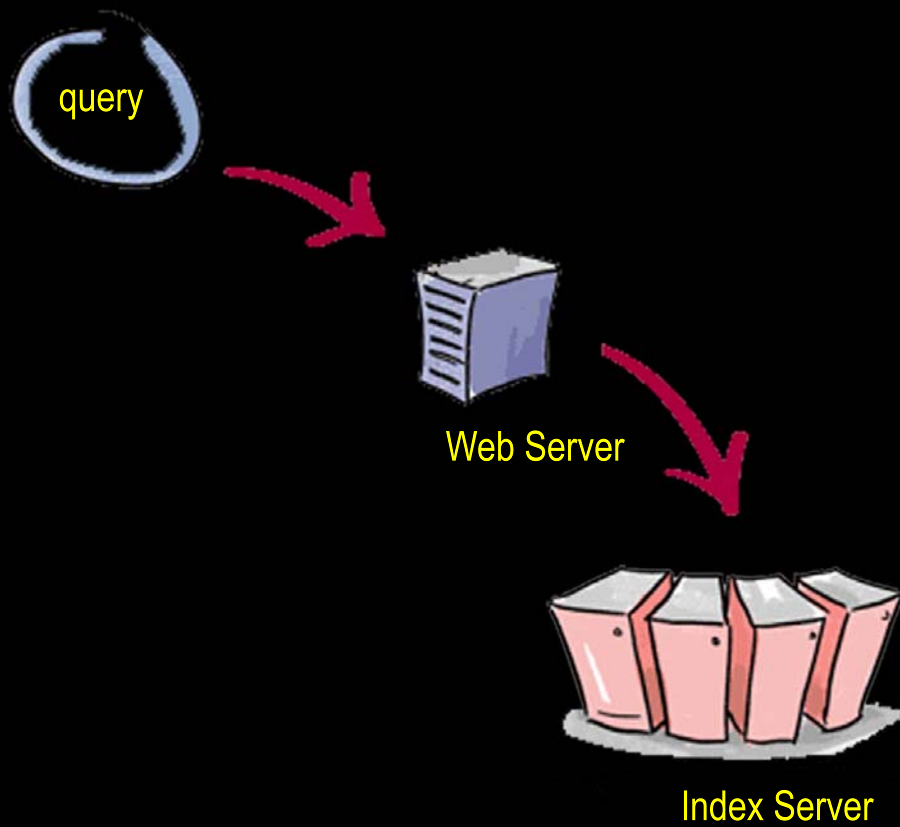
- Measure the importance of each page
- The measure should be Independent of any query
- Compute these measures off-line before queries are processed
- Google's PageRank© distinguishes it from all competitors
  - — Primarily determined by the link structure of the Web
  - — Tempered by some content considerations

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

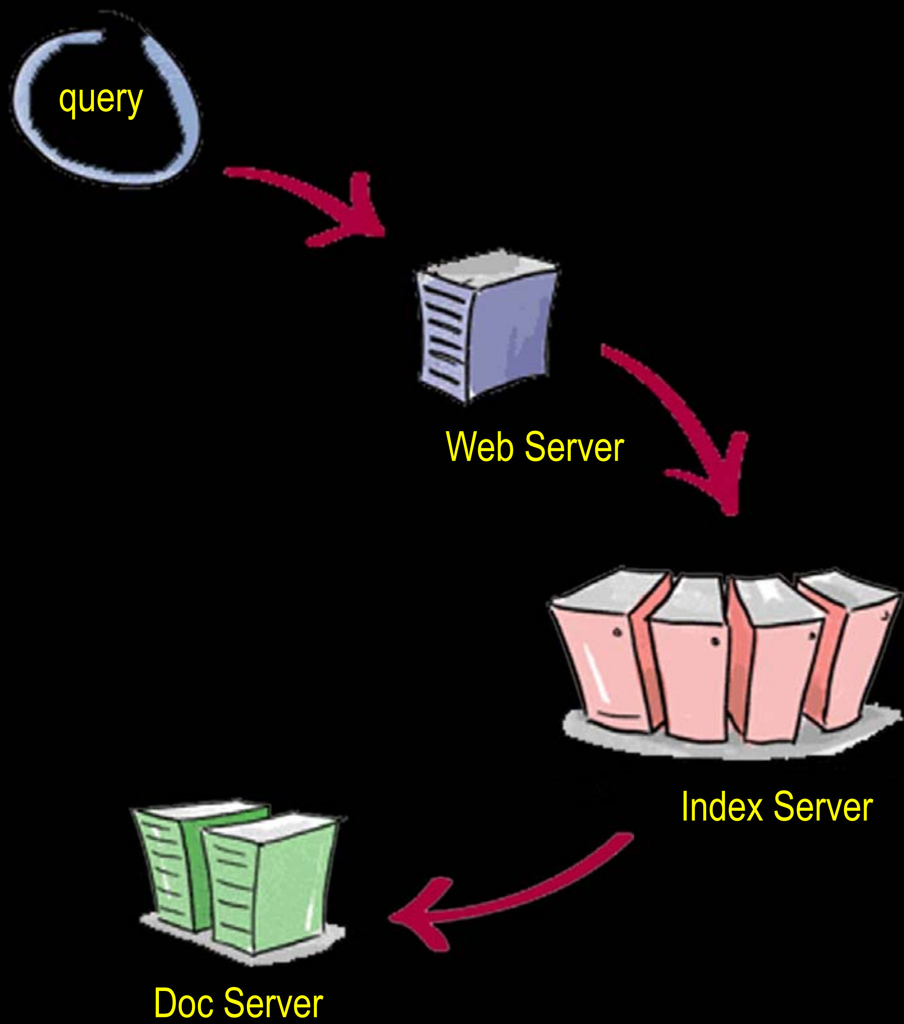
The Process



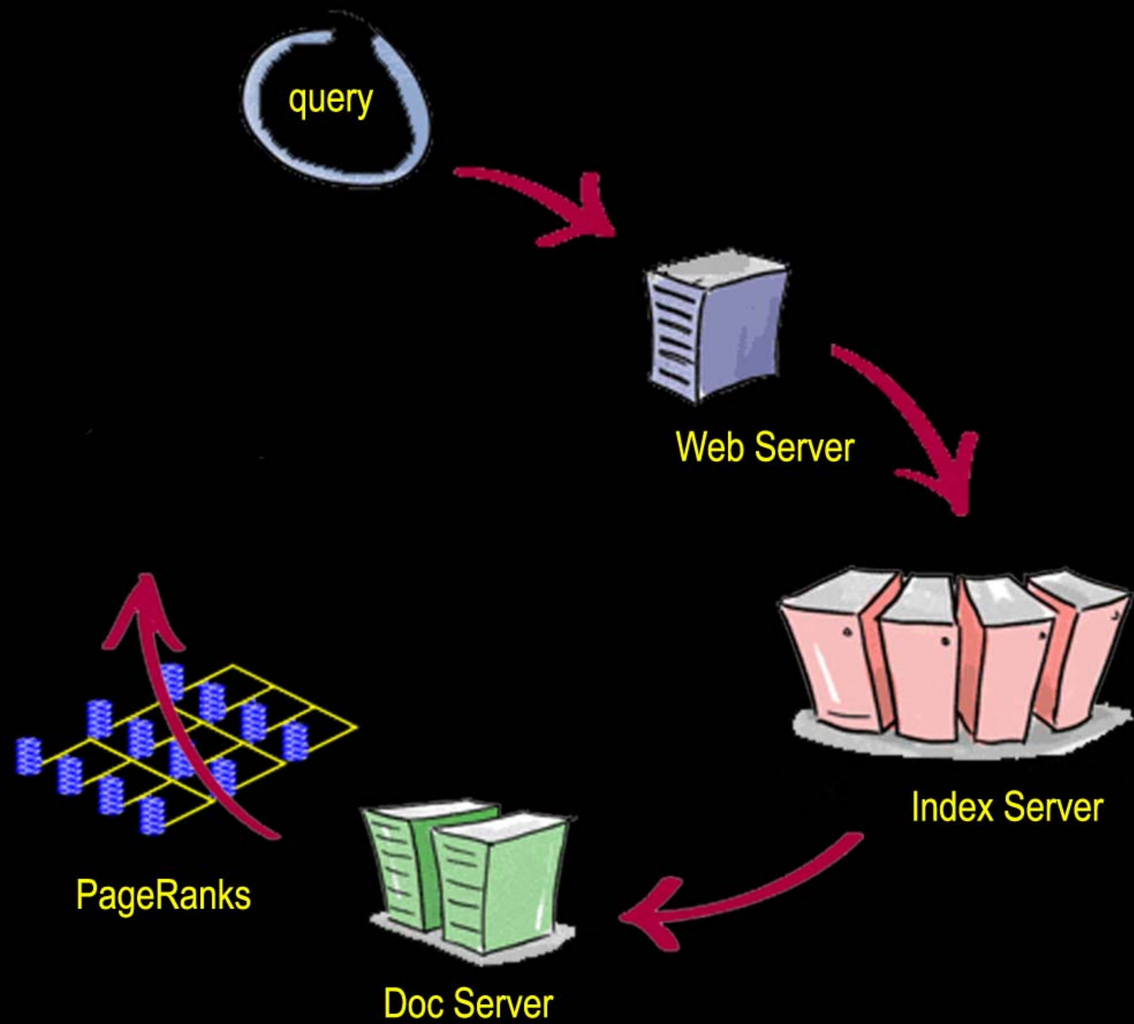
The Process



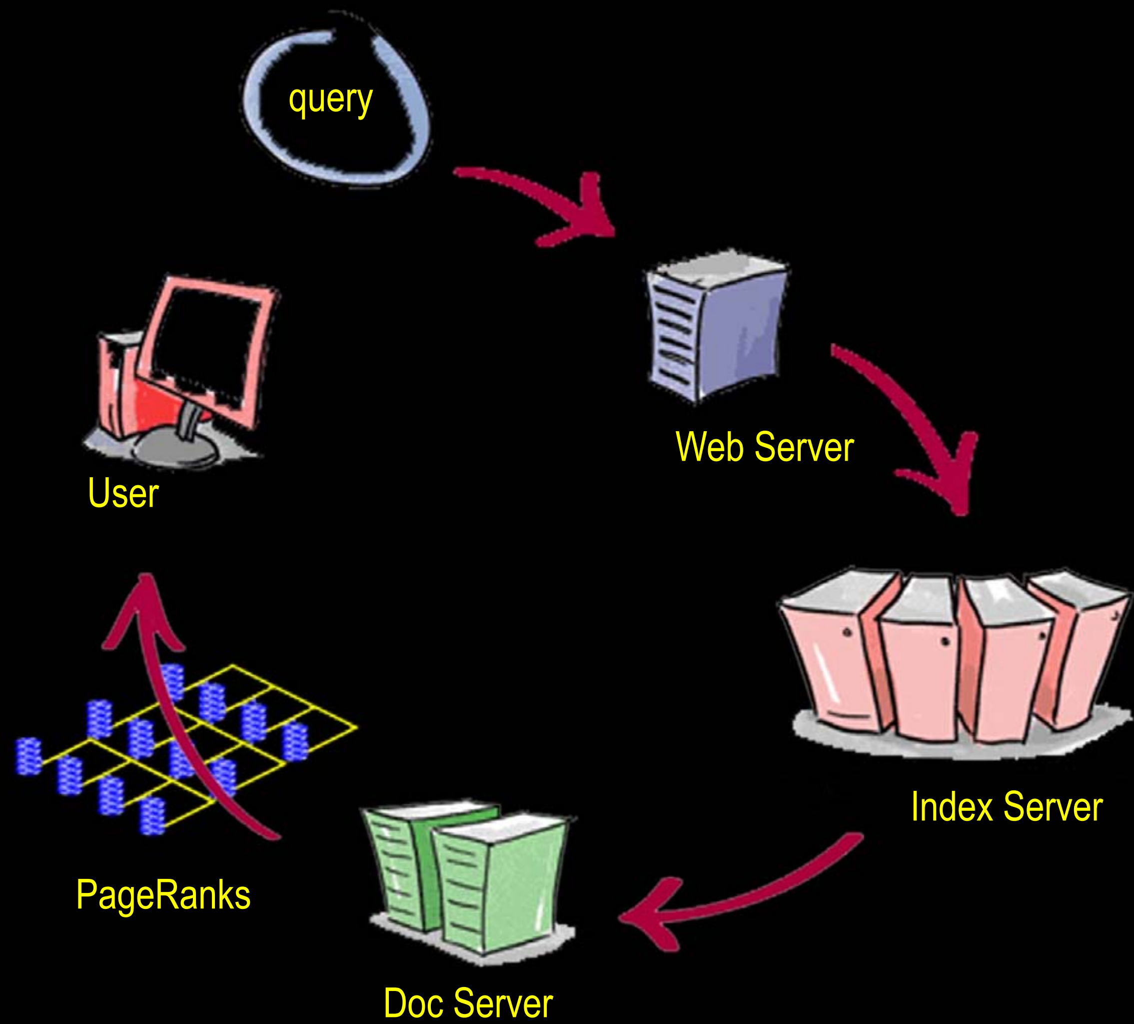
The Process



The Process



The Process



Search

1.8 million

Google shares given to Stanford University for an exclusive license of the PageRank patent (owned by the university). They were sold in 2005 for **$336 million.**

Daily page views for Google.com

7.2 billion

Monthly worldwide searches on Google sites

87.8 billion

Global search marketshare

85.78%

Daily visitors to Google.com

620 million

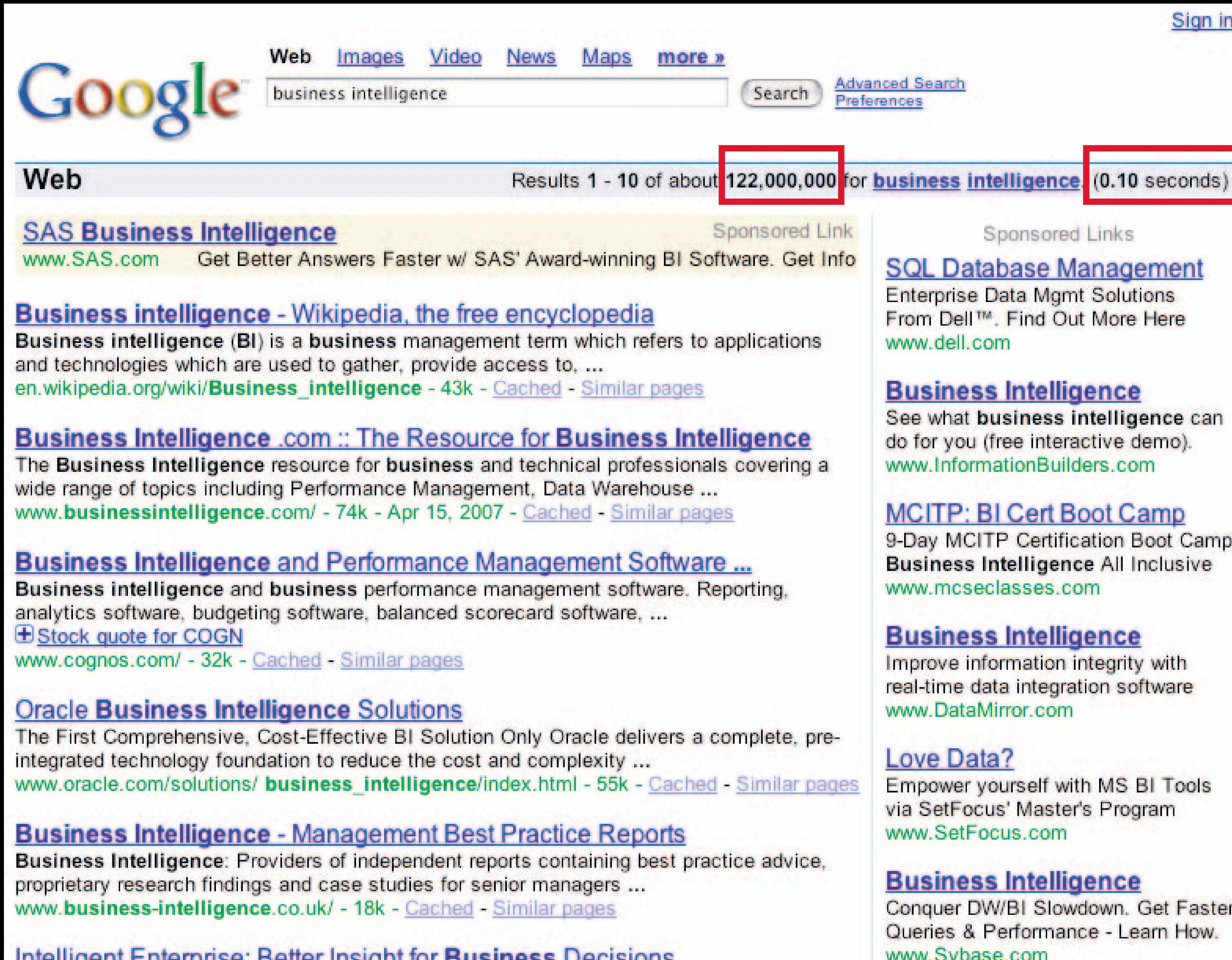
Google.com's global website ranking

1

The amount of data processed daily by Google

20 PB

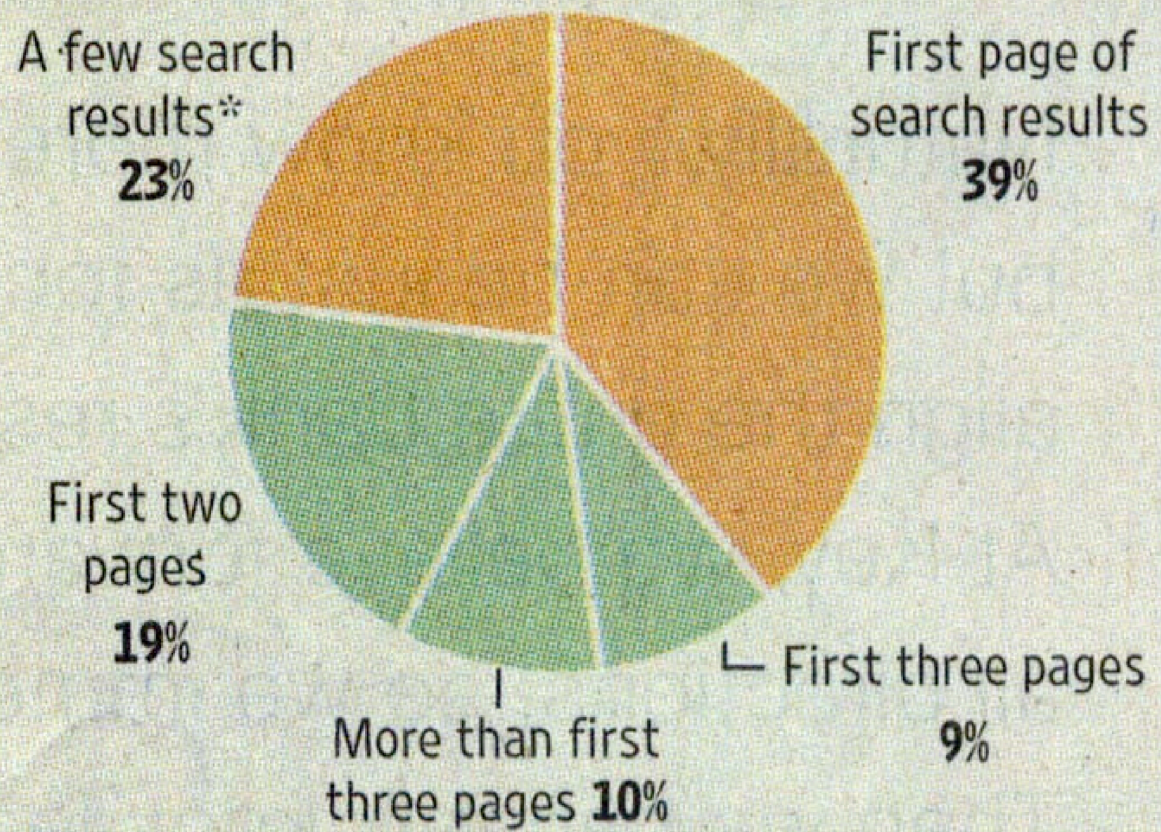
Google Search support for fictional languages:
Leetspeak (H4x0r), Klingon, Pig Latin, Elmer Fudd and Bork, bork, bork!


Sign in
Web Images Video News Maps more »
Google
business intelligence
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)
SAS Business Intelligence
Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info
Business intelligence - Wikipedia, the free encyclopedia
Business intelligence (BI) is a business management term which refers to applications and technologies which are used to gather, provide access to, ...
en.wikipedia.org/wiki/Business_intelligence - 43k - Cached - Similar pages
Business Intelligence .com :: The Resource for Business Intelligence
The Business Intelligence resource for business and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ...
www.businessintelligence.com/ - 74k - Apr 15, 2007 - Cached - Similar pages
Business Intelligence and Performance Management Software ...
Business intelligence and business performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ...
Stock quote for COGN
www.cognos.com/ - 32k - Cached - Similar pages
Oracle Business Intelligence Solutions
The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ...
www.oracle.com/solutions/ business_intelligence/index.html - 55k - Cached - Similar pages
Business Intelligence - Management Best Practice Reports
Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ...
www.business-intelligence.co.uk/ - 18k - Cached - Similar pages
Intelligent Enterprise: Better Insight for Business Decisions
Sponsored Links
SQL Database Management
Enterprise Data Mgmt Solutions From Dell™. Find Out More Here
www.dell.com
Business Intelligence
See what business intelligence can do for you (free interactive demo).
www.InformationBuilders.com
MCITP: BI Cert Boot Camp
9-Day MCITP Certification Boot Camp Business Intelligence All Inclusive
www.mcseclasses.com
Business Intelligence
Improve information integrity with real-time data integration software
www.DataMirror.com
Love Data?
Empower yourself with MS BI Tools via SetFocus' Master's Program
www.SetFocus.com
Business Intelligence
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com

The Wall Street Journal, April 13, 2007

Take Your Pick

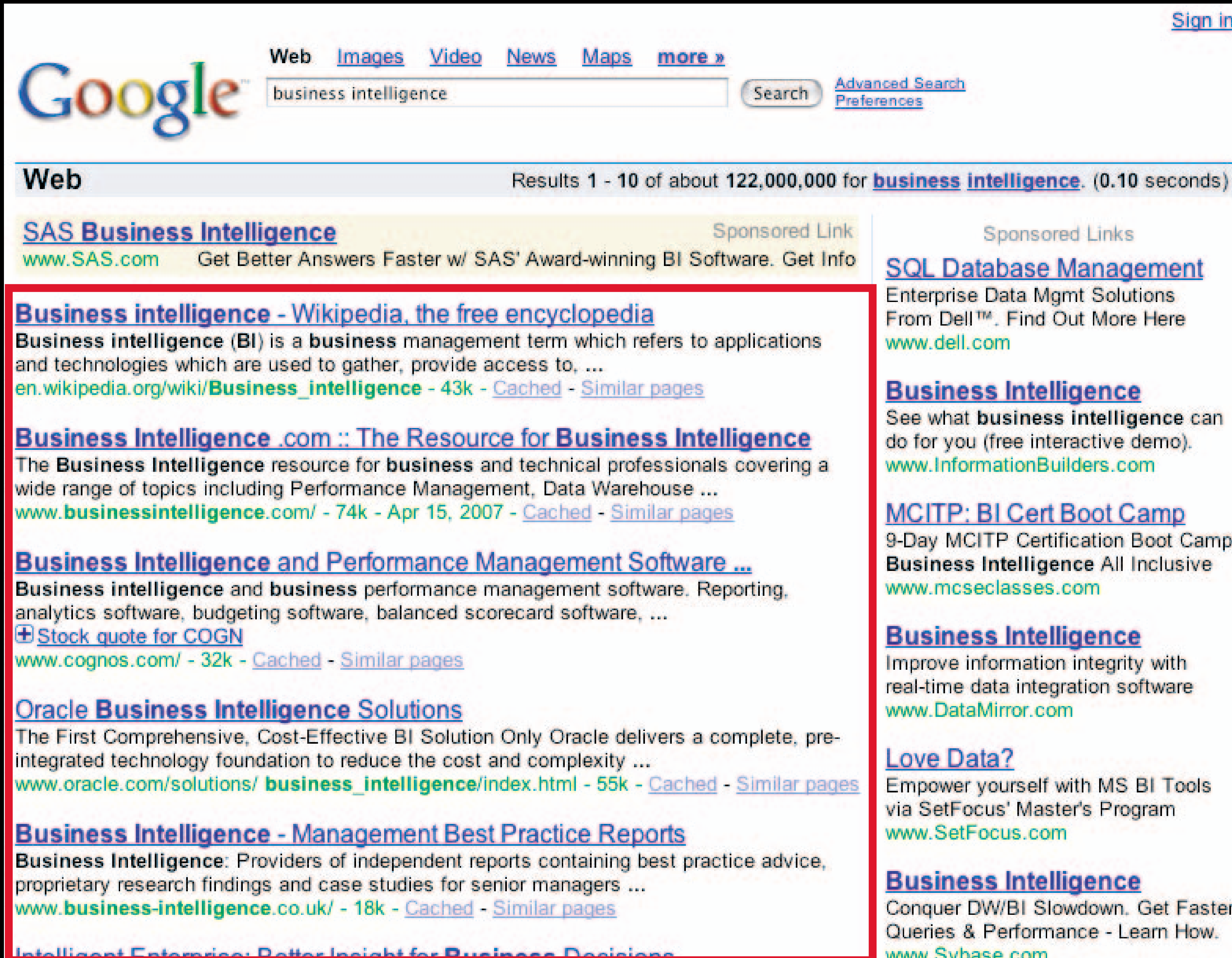
Amount of Internet search results that Web surfers typically scan before selecting one

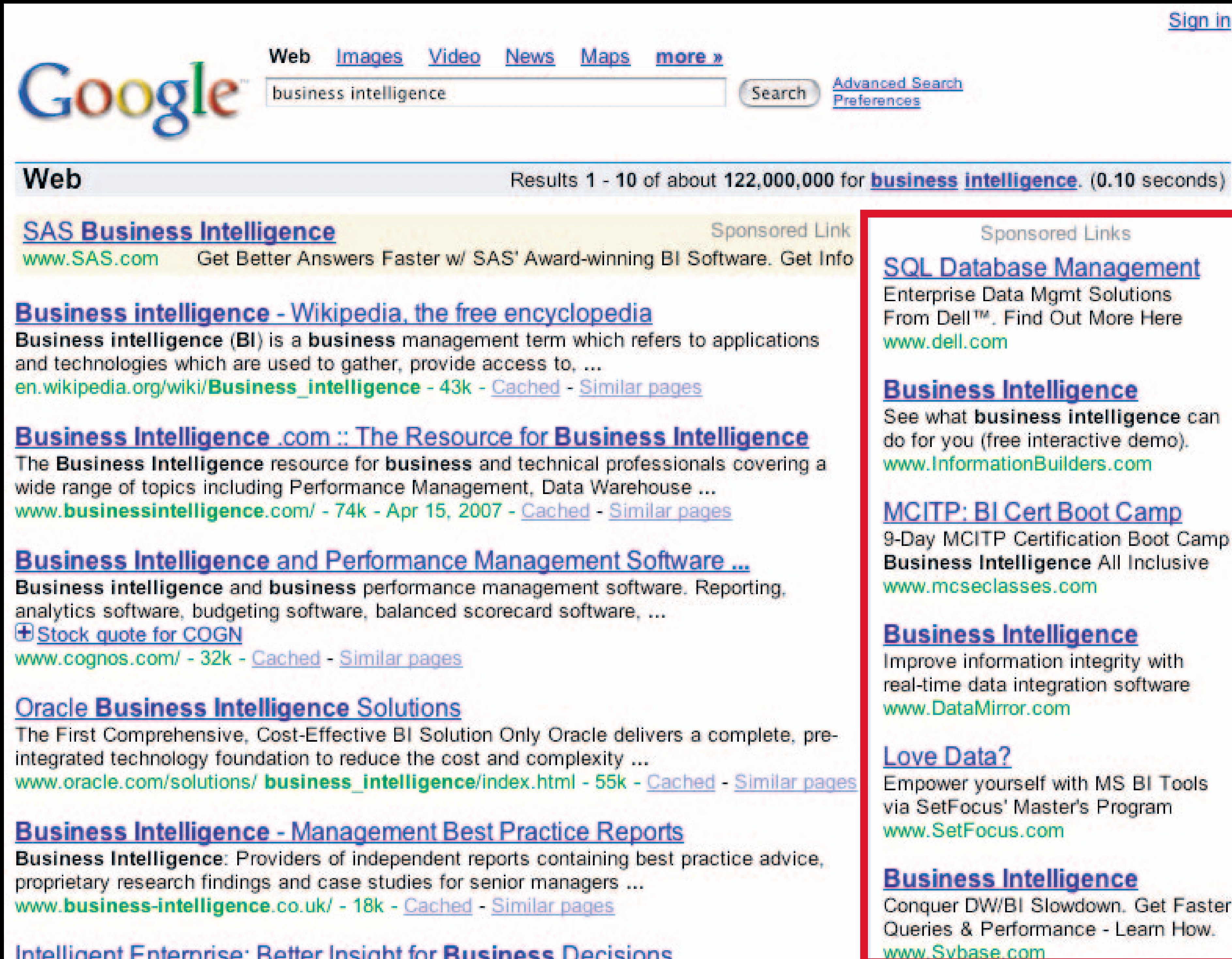


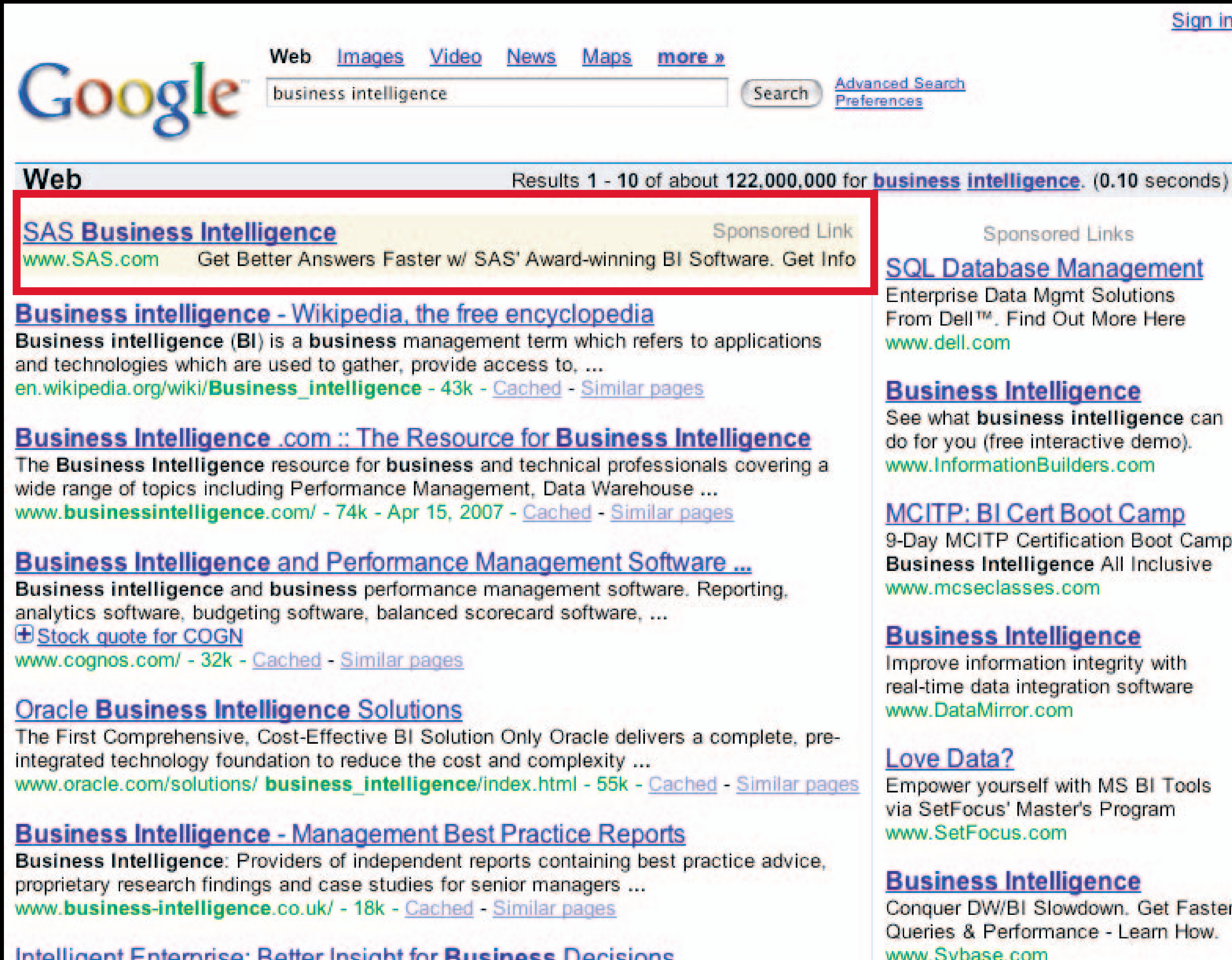
*Top results without reading through the whole page

Note: Sample size is 2,369 people

Sources: JupiterResearch; iProspect


Sign in
Web Images Video News Maps more »
Google
business intelligence
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)
SAS Business Intelligence
Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info
Business intelligence - Wikipedia, the free encyclopedia
Business intelligence (BI) is a business management term which refers to applications and technologies which are used to gather, provide access to, ...
en.wikipedia.org/wiki/Business_intelligence - 43k - Cached - Similar pages
Business Intelligence .com :: The Resource for Business Intelligence
The Business Intelligence resource for business and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ...
www.businessintelligence.com/ - 74k - Apr 15, 2007 - Cached - Similar pages
Business Intelligence and Performance Management Software ...
Business intelligence and business performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ...
Stock quote for COGN
www.cognos.com/ - 32k - Cached - Similar pages
Oracle Business Intelligence Solutions
The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ...
www.oracle.com/solutions/ business_intelligence/index.html - 55k - Cached - Similar pages
Business Intelligence - Management Best Practice Reports
Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ...
www.business-intelligence.co.uk/ - 18k - Cached - Similar pages
Sponsored Links
SQL Database Management
Enterprise Data Mgmt Solutions From Dell™. Find Out More Here
www.dell.com
Business Intelligence
See what business intelligence can do for you (free interactive demo).
www.InformationBuilders.com
MCITP: BI Cert Boot Camp
9-Day MCITP Certification Boot Camp Business Intelligence All Inclusive
www.mcseclasses.com
Business Intelligence
Improve information integrity with real-time data integration software
www.DataMirror.com
Love Data?
Empower yourself with MS BI Tools via SetFocus' Master's Program
www.SetFocus.com
Business Intelligence
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com


Sign in
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

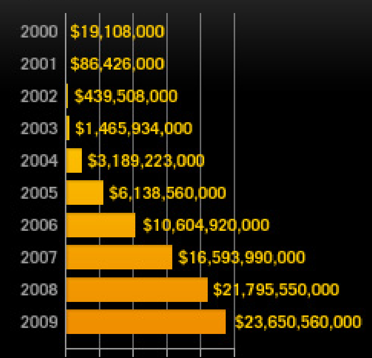

Sign in
Web Images Video News Maps more »
business intelligence
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)
SAS Business Intelligence
Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info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

Money

$2,718,281,828

The target for Google's IPO on April 29, 2004. This somewhat strange number is the equivalent of the mathematical constant e in billions (**e ≈ 2.718281828**).

Revenue

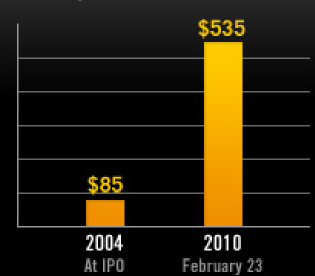


Profit

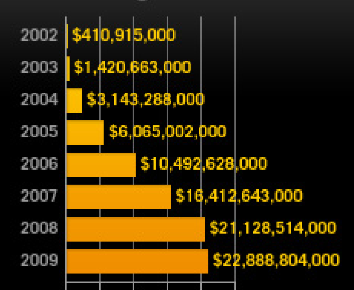
Percent of revenue from advertising

97%

Stock price



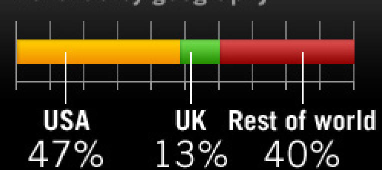
Advertising revenue



Selected acquisitions

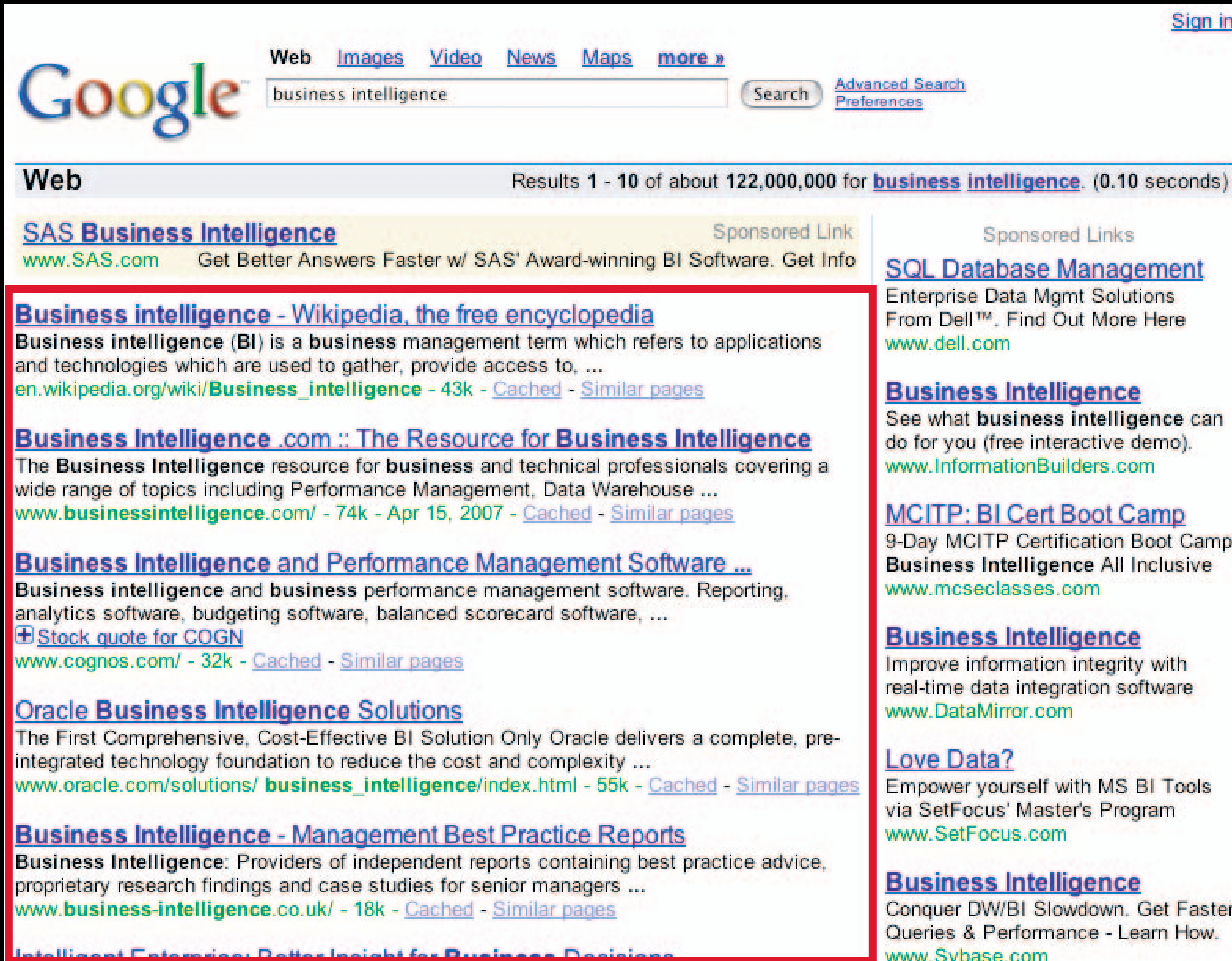
Date	Company	Type
Feb, 2003	Pyra Labs	Weblog software
Mar, 2005	Urchin	Web analytics
Aug 17, 2005	Android	Mobile software
Oct 9, 2006	YouTube	Video sharing
Apr 13, 2007	DoubleClick	Online advertising
July 9, 2007	Postini	Email security
Nov 9, 2009	AdMob	Mobile advertising

Revenue by geography



Assets (Dec 31, 2009)

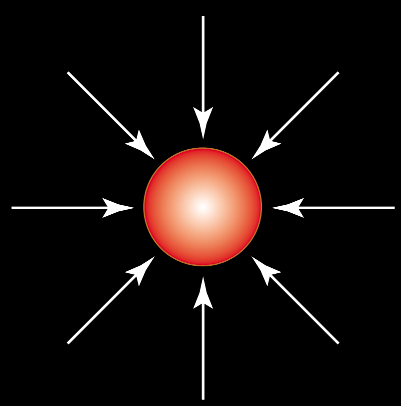
$40.5 billion



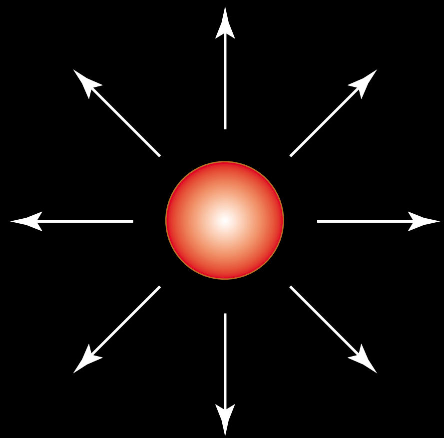
Sign in
Web Images Video News Maps more »
Google
business intelligence
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 122,000,000 for business intelligence. (0.10 seconds)
SAS Business Intelligence
Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info
Business intelligence - Wikipedia, the free encyclopedia
Business intelligence (BI) is a business management term which refers to applications and technologies which are used to gather, provide access to, ...
en.wikipedia.org/wiki/Business_intelligence - 43k - Cached - Similar pages
Business Intelligence .com :: The Resource for Business Intelligence
The Business Intelligence resource for business and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ...
www.businessintelligence.com/ - 74k - Apr 15, 2007 - Cached - Similar pages
Business Intelligence and Performance Management Software ...
Business intelligence and business performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ...
Stock quote for COGN
www.cognos.com/ - 32k - Cached - Similar pages
Oracle Business Intelligence Solutions
The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ...
www.oracle.com/solutions/ business_intelligence/index.html - 55k - Cached - Similar pages
Business Intelligence - Management Best Practice Reports
Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ...
www.business-intelligence.co.uk/ - 18k - Cached - Similar pages
Sponsored Links
SQL Database Management
Enterprise Data Mgmt Solutions From Dell™. Find Out More Here
www.dell.com
Business Intelligence
See what business intelligence can do for you (free interactive demo).
www.InformationBuilders.com
MCITP: BI Cert Boot Camp
9-Day MCITP Certification Boot Camp Business Intelligence All Inclusive
www.mcseclasses.com
Business Intelligence
Improve information integrity with real-time data integration software
www.DataMirror.com
Love Data?
Empower yourself with MS BI Tools via SetFocus' Master's Program
www.SetFocus.com
Business Intelligence
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com

How To Measure “Importance”

Landmark Result Paper

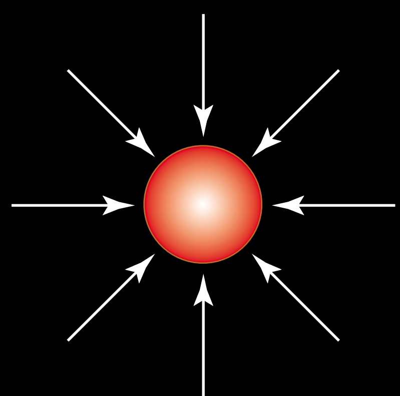


Survey Paper—Big Bib



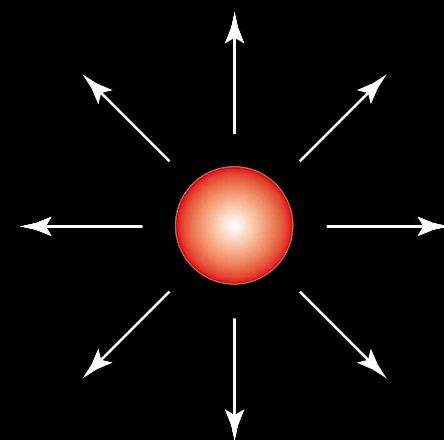
How To Measure "Importance"

Landmark Result Paper



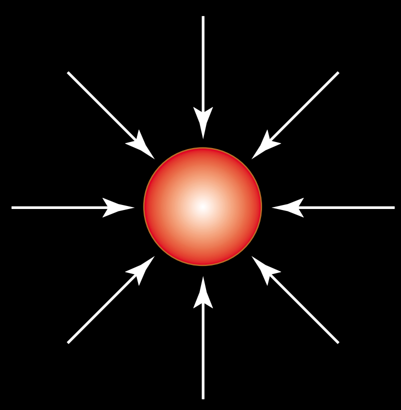
Authorities

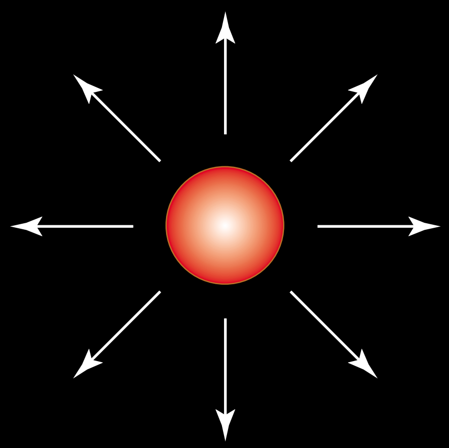
Survey Paper—Big Bib



Hubs

How To Measure "Importance"





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

HITS

Hypertext Induced Topic Search (1998)

Jon Kleinberg

Determine Authority & Hub Scores

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

HITS

Hypertext Induced Topic Search (1998)

Jon Kleinberg

Determine Authority & Hub Scores

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

Successive Refinement

- Start with $h_i = 1$ for all pages P_i $\Rightarrow$ $\mathbf{h}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

HITS

Hypertext Induced Topic Search (1998)

Jon Kleinberg

Determine Authority & Hub Scores

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

Successive Refinement

- Start with $h_i = 1$ for all pages P_i $\Rightarrow$ $\mathbf{h}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
- Define Authority Scores (on the first pass)

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

HITS

Hypertext Induced Topic Search (1998)

Jon Kleinberg

Determine Authority & Hub Scores

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

Successive Refinement

- Start with $h_i = 1$ for all pages $P_i \quad \Rightarrow \quad \mathbf{h}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
- 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

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

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

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

$\ddots$

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

$\ddots$

Combined Iterations

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

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

$\ddots$

Combined Iterations

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

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

$\ddots$

Combined Iterations

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

HITS Algorithm

Refine Hub Scores

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

Successively Re-refine Authority & Hub Scores

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

$$\ddots$$

Combined Iterations

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

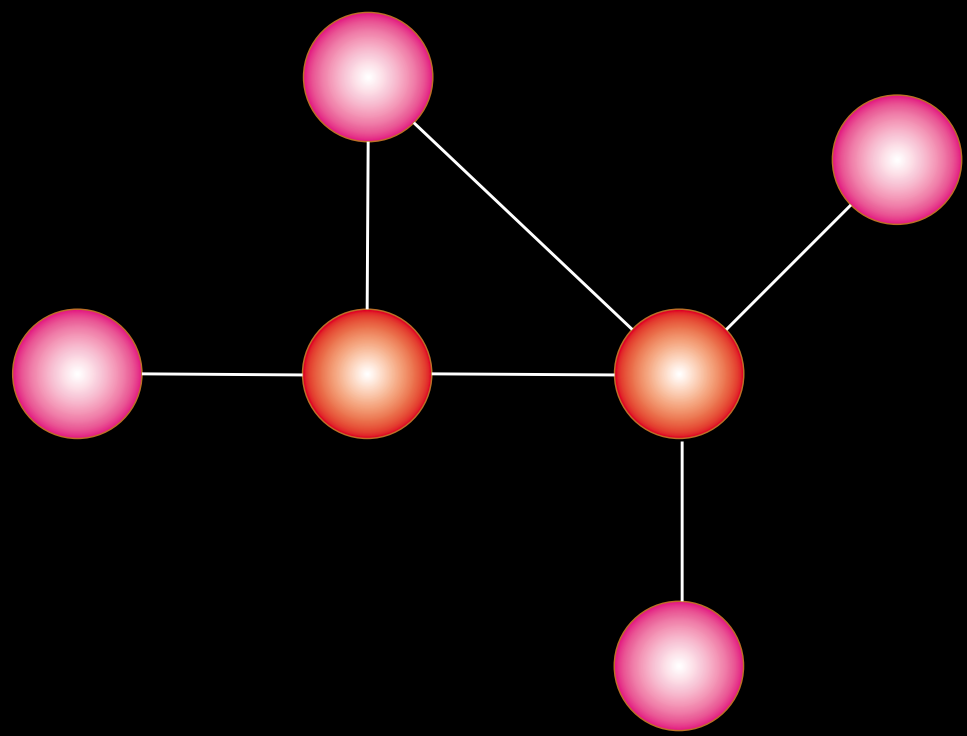
!! **A lot of work** !!

Compromise

1. Do direct query matching

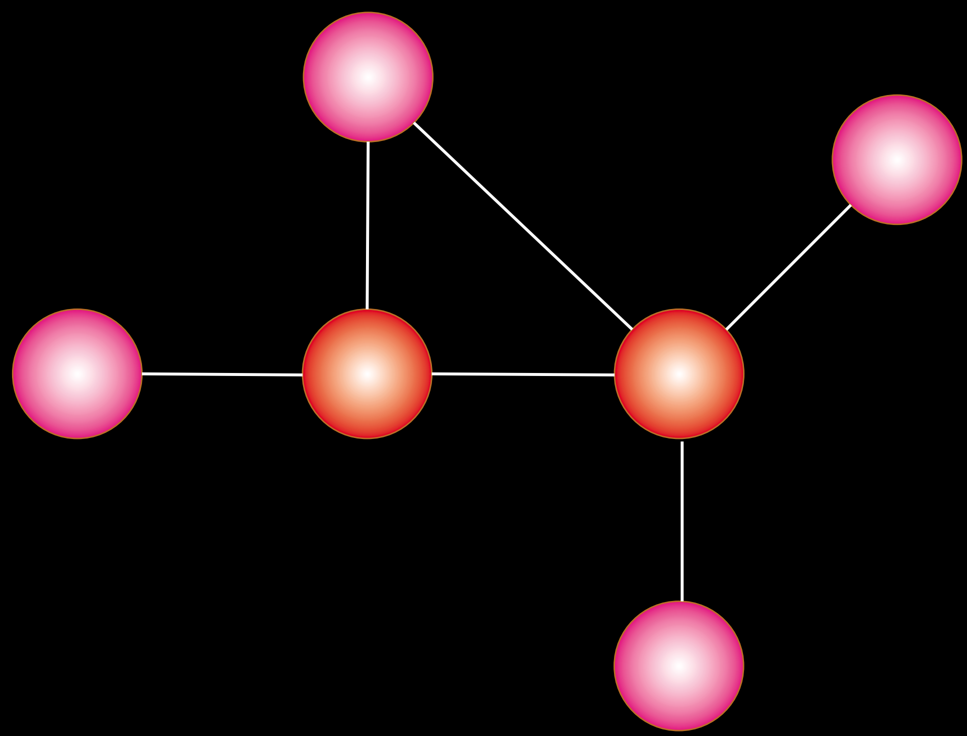
Compromise

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

Disadvantages

- Too much has to happen while client is waiting

Pros & Cons

Advantages

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

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

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

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™

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 rolling 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 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.

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

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.

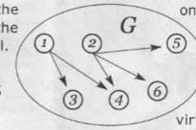
So what's the solution to search? While computer science was trying to coax an answer from its collective hard drive, it was sitting right there in the stacks all along. Who could have guessed that when Eugene Garfield went all bibliometric and devised a system to find out how much a journal mattered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

$$\boxed{\begin{aligned} \bar{h} &= \lambda A \bar{a} \\ \bar{a} &= \lambda A' \bar{h} \end{aligned}}$$

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.



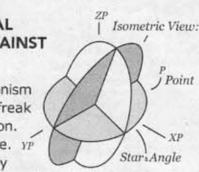
THE ALGORITHM IS RELATIVELY SIMPLE, IF YOU'RE SOME KIND OF SAVANT.

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 $a(p)$ → $\boxed{a = A^T h}$ and a non-negative hub weight $h(p)$. → $\boxed{h = Aa}$ 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.

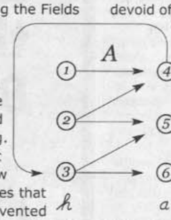
IT'S A GOOD THING ROBERT FROST NEVER WROTE AN ALGORITHM.

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

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

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

Google's PageRank

(Lawrence Page & Sergey Brin 1998)

The Google Goals

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

PageRank

Google's Original Idea

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

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

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

PageRank

Google's Original Idea

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

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

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

Successive Refinement

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

PageRank

Google's Original Idea

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

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

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

Successive Refinement

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

Iteratively refine rankings for each page

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

PageRank

Google's Original Idea

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

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

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

Successive Refinement

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

Iteratively refine rankings for each page

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

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

PageRank

Google's Original Idea

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

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

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

Successive Refinement

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

Iteratively refine rankings for each page

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

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

$$\ddots$$

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

In Matrix Notation

After Step k

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

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

In Matrix Notation

After Step k

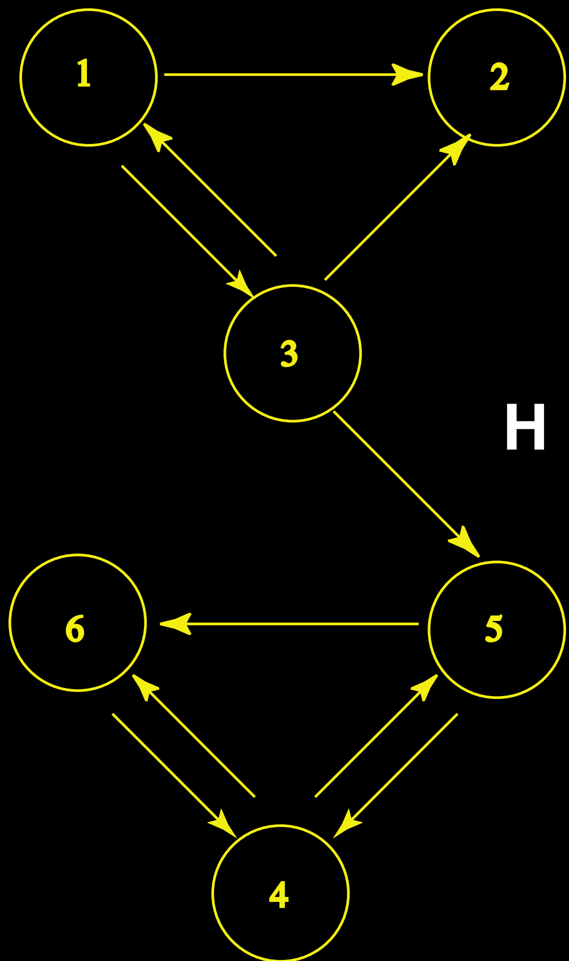
— $\boldsymbol{\pi}_k^T = [r_k(P_1), r_k(P_2), \cdots, r_k(P_n)]$

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

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

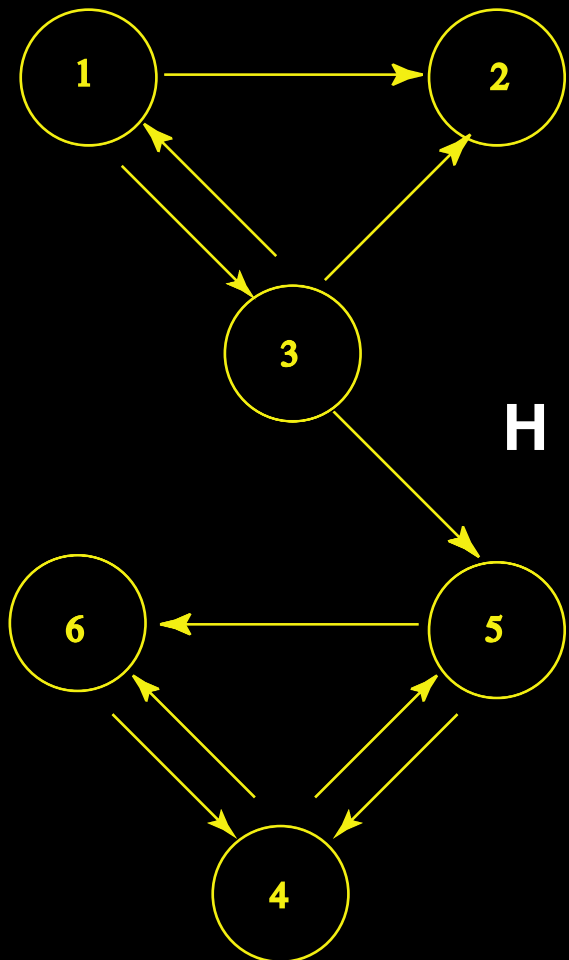
$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{H}$$

Tiny Web



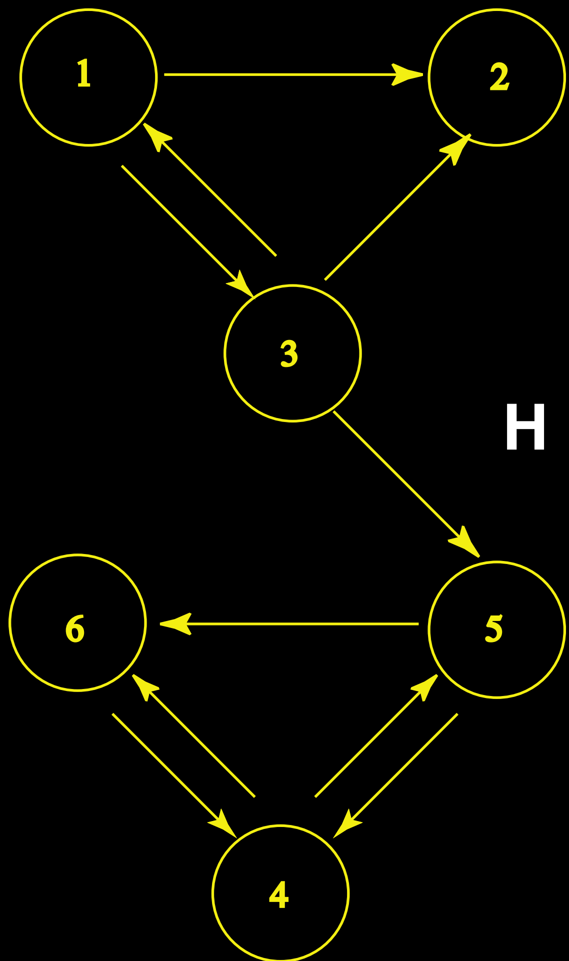
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \left(\begin{array}{cccccc} \; & \; & \; & \; & \; & \; \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{array}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

Tiny Web



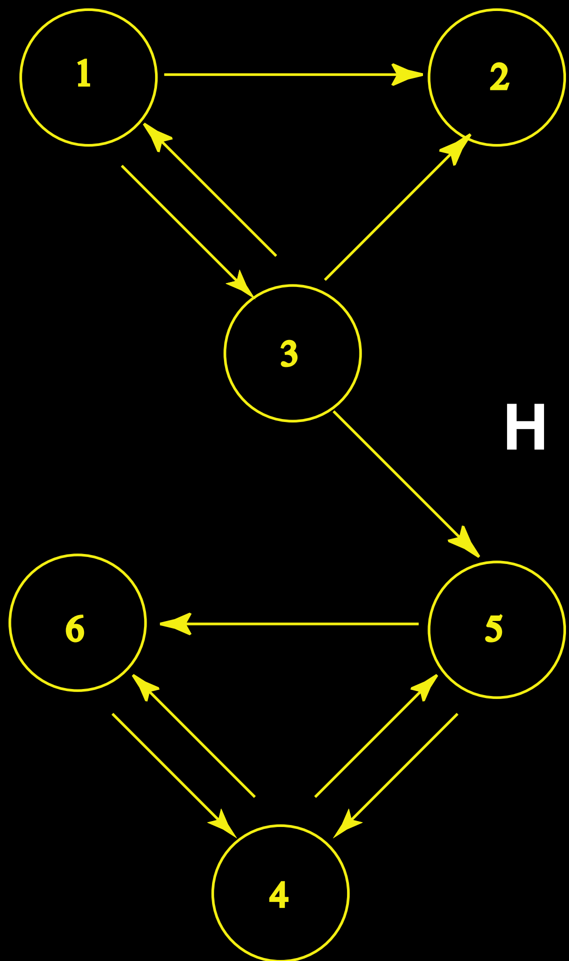
$$\mathbf{H} = \begin{array}{c c} & \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{array}{c} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} & \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



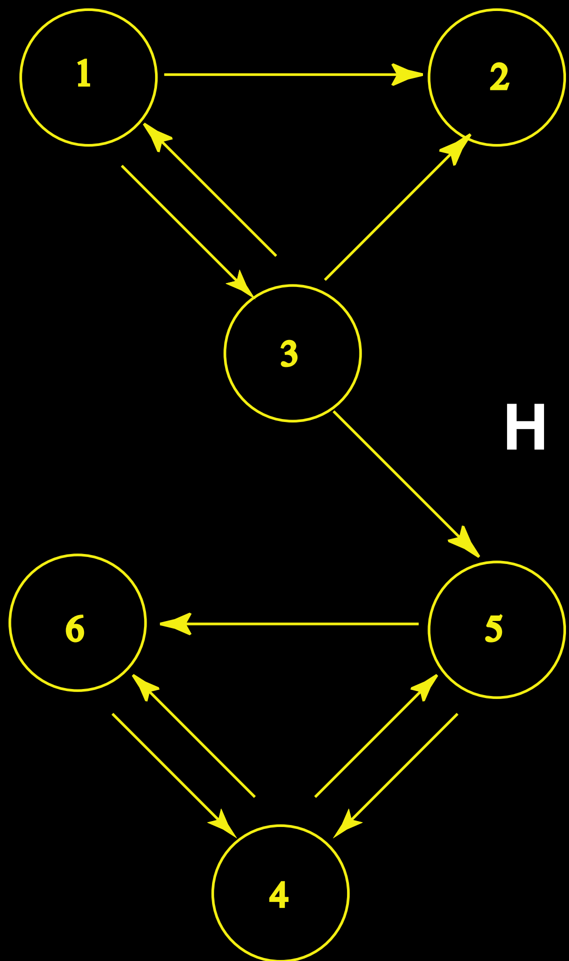
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix}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

Tiny Web



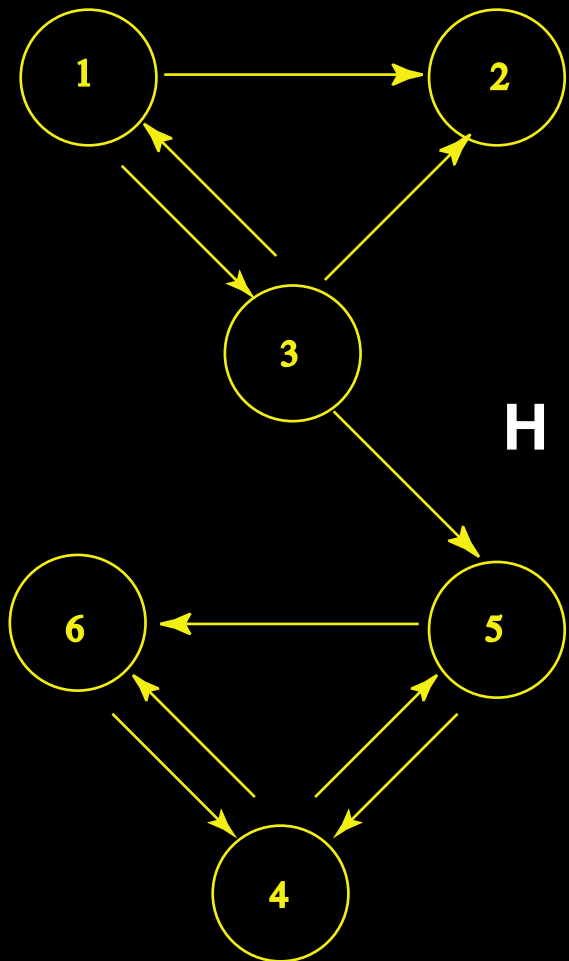
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ & & & & & \\ & & & & & \\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

Tiny Web



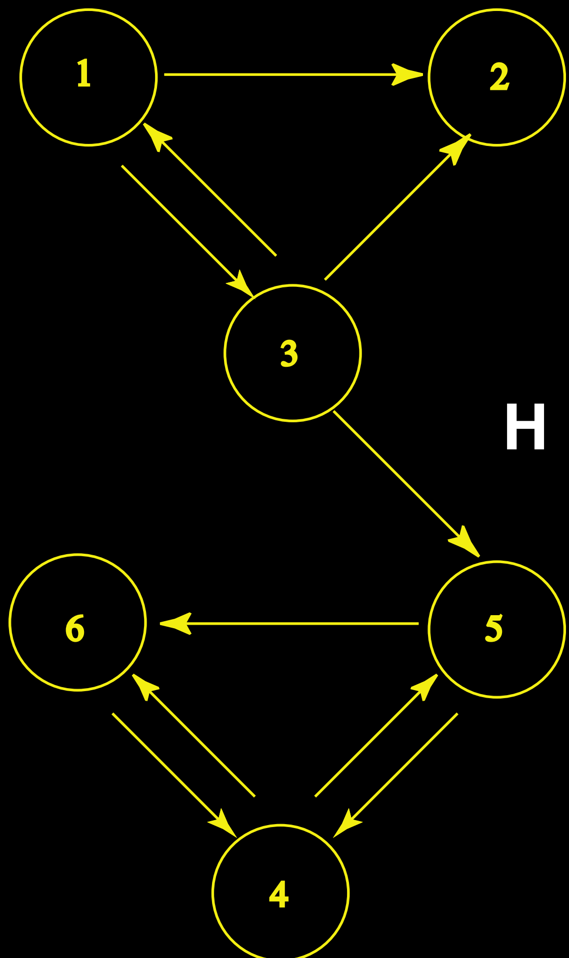
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \left(\begin{array}{cccccc} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0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

Tiny Web



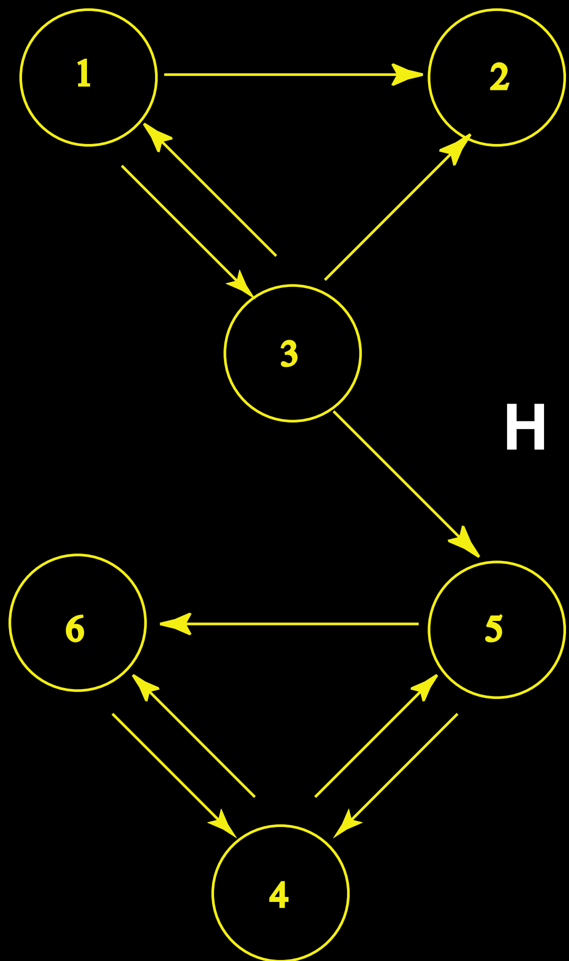
$$
\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array}
\begin{array}{c}
\begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\
\begin{pmatrix}
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2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

Tiny Web



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

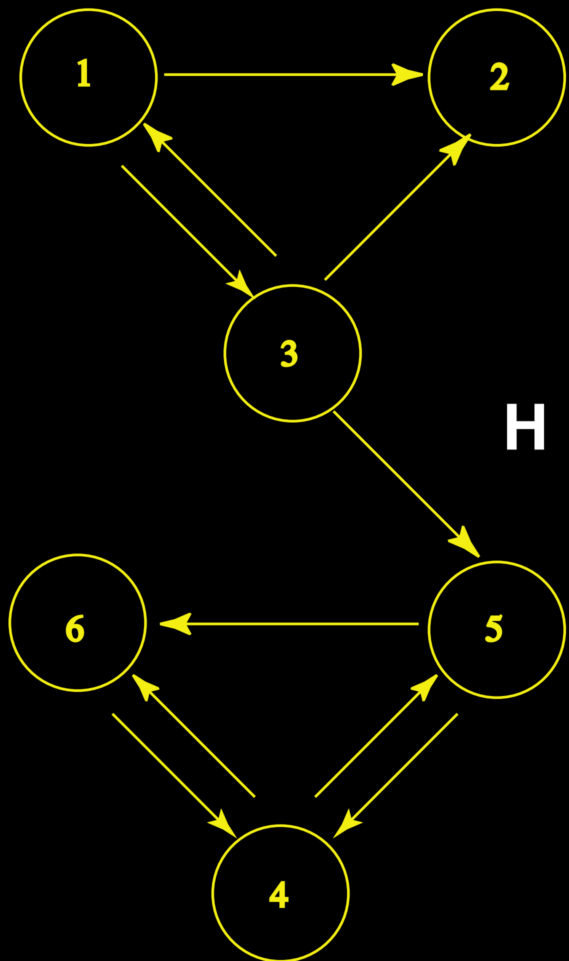
Tiny Web



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

- A random walk on the Web Graph

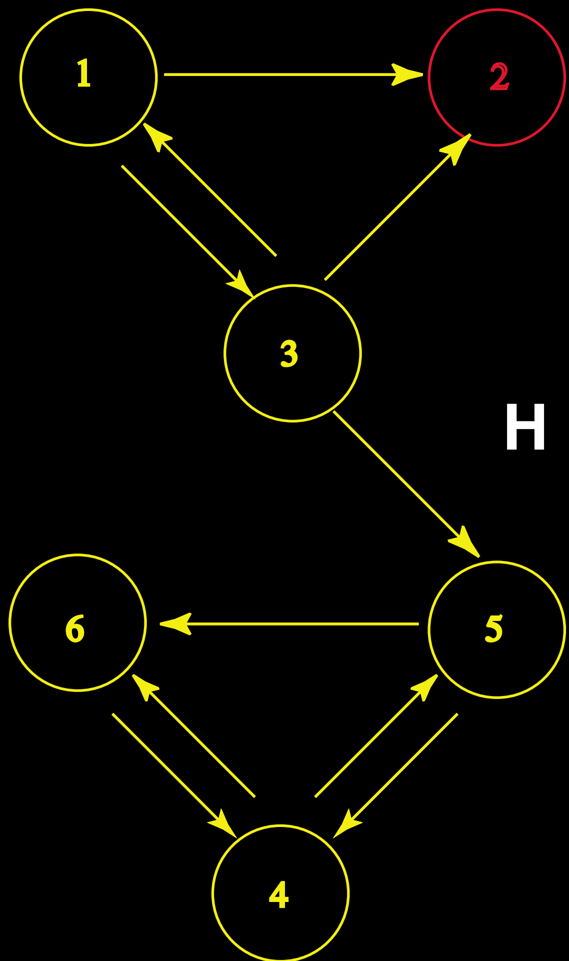
Tiny Web



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

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

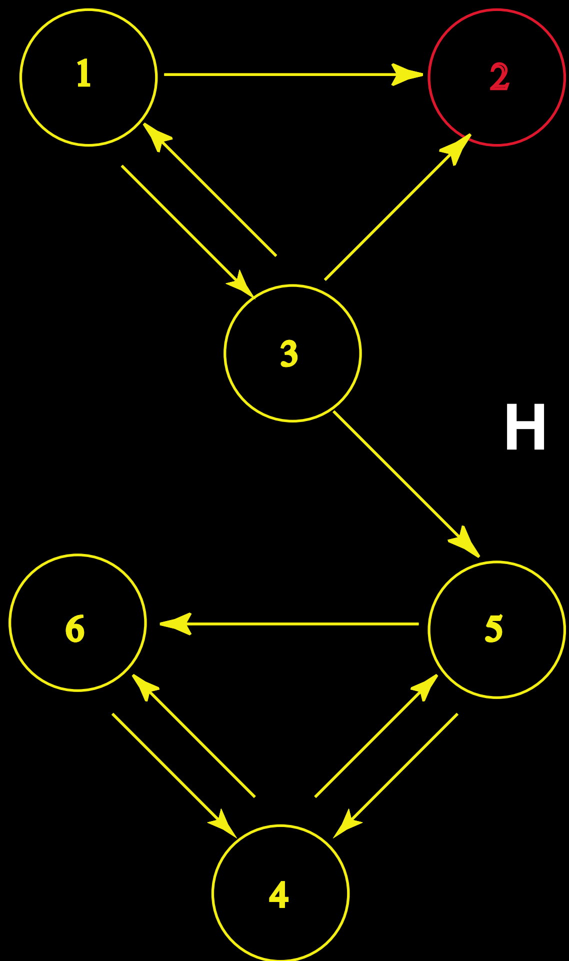
Tiny Web



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

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

Tiny Web



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

- ▷ A random walk on the Web Graph
- ▷ PageRank = π_i = amount of time spent at P_i
- ▷ Dead end page (nothing to click on) — a “dangling node”
- ▷ $\pi^T = (0, 1, 0, 0, 0, 0) \Longrightarrow$ Page P_2 is a “rank sink”

The Fix

Allow Web Surfers To Make Random Jumps

The Fix

Allow Web Surfers To Make Random Jumps

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

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

Nasty Problem

The Web Graph Is Not Strongly Connected

Nasty Problem

The Web Graph Is Not Strongly Connected

— i.e., S is a reducible matrix

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

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle

$$P_i \longrightarrow P_j \longrightarrow P_i$$

$$\begin{array}{c} \boldsymbol{\pi}^T = \left(\mathbf{0} \quad \cdots \quad \mathbf{1/2} \quad \mathbf{0} \quad \cdots \quad \mathbf{1/2} \quad \mathbf{0} \quad \cdots \quad \mathbf{0} \right) \end{array}$$

$$\begin{array}{ccccccccccc} \phantom{\boldsymbol{\pi}^T = (0} & & \uparrow & & & \uparrow & & & \\ & & i & & & j & & & \end{array}$$

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad 0 < \alpha < 1 \qquad (\alpha \approx .85)$

$$\mathbf{E} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \qquad \boldsymbol{\pi}^T\mathbf{G} = \boldsymbol{\pi}^T$$

The Google Fix

Allow A Random Jump From Any Page

— $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad 0 < \alpha < 1$ $\quad (\alpha \approx .85)$

$$\mathbf{E} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \qquad \boldsymbol{\pi}^T\mathbf{G} = \boldsymbol{\pi}^T$$

$$\mathbf{E} = \mathbf{u}\mathbf{v}^T = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & \cdots & u_1v_n \\ u_2v_1 & u_2v_2 & \cdots & u_2v_n \\ \vdots & \vdots & \ddots & \vdots \\ u_nv_1 & u_nv_2 & \cdots & u_nv_n \end{bmatrix}$$

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

Always Changing

PR Augmented With Content Scores For Final Rankings

“Metrics” Are Proprietary — But Known Examples

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

Always Changing

PR Augmented With Content Scores For Final Rankings

"Metrics" Are Proprietary — But Known Examples

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

Every Thursday

Wired March 2010 http://www.wired.com

- Three dozen engineers, product managers, and executives make Google smarter
- This year (2010), Google plans to introduce about 550 improvements

Improvement History

- Backrub [September 1997]
  - — Had run on Stanford servers for almost two years—renamed Google.
- New algorithm [August 2001]
  - — Search algorithm completely revamped—incorporated additional ranking criteria
- Local connectivity analysis [February 2003]
  - — Gives more weight to links from authoritative sites
- Fritz [Summer 2003]
  - — Update the index constantly instead of in big batches
- Personalization [June 2005]
  - — Mine search behavior to provide individualized results
- Bigdaddy [December 2005]
  - — Engine update allows for more-comprehensive Web crawling
- Universal search [May 2007]
  - — Provide links to any medium (image, news, books) on the same results page
- Real-Time Search [December 2009]
  - — Results from Twitter and blogs as they are published

Conclusion

Google and PageRank ~~Is~~ Chang~~ing~~ed The World

Thank You