Computing and Information Retrieval

The Big Picture

Amy N. Langville

Department of Mathematics
College of Charleston
Charleston, SC

DOE-CSGF Meeting 6/21/05

Outline

- Introduction to Information Retrieval (IR)
- Traditional and Web Search
- Problems in Web Search
- Innovations

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	(trad. search)
1989	Berner-Lee's WWW	(web search)

Traditional Search

Two Primary Goals:

- Clustering documents
- Processing user queries
  - find similar documents
  - find similar terms

Vector Space Model (1960s and 1970s)

Gerard Salton's Information Retrieval System

SMART: System for the Mechanical Analysis and Retrieval of Text

(Salton's Magical Automatic Retriever of Text)

- turn n textual documents into n document vectors $\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n$
- create term-by-document matrix $\mathbf{A}_{m \times n} = [\, \mathbf{d}_1 | \mathbf{d}_2 | \cdots | \mathbf{d}_n \,]$
- to retrieve info., create query vector $\mathbf{q}$, which is a pseudo-doc

GOAL: find doc. $\mathbf{d}_i$ closest to $\mathbf{q}$

— angular cosine measure used: $\delta_i = cos\, \theta_i = \mathbf{q}^T \mathbf{d}_i / (\|\mathbf{q}\|_2 \|\mathbf{d}_i\|_2)$

Example from Berry's book

Terms

T1: Bab(y,ies,y's)

T2: Child(ren's)

T3: Guide

T4: Health

T5: Home

T6: Infant

T7: Proofing

T8: Safety

T9: Toddler

Documents

D1: Infant & Toddler First Aid

D2: Babies & Children's Room (For Your Home)

D3: Child Safety at Home

D4: Your Baby's Health & Safety : From Infant to Toddler

D5: Baby Proofing Basics

D6: Your Guide to Easy Rust Proofing

D7: Beanie Babies Collector's Guide

Example from Berry's book

Terms

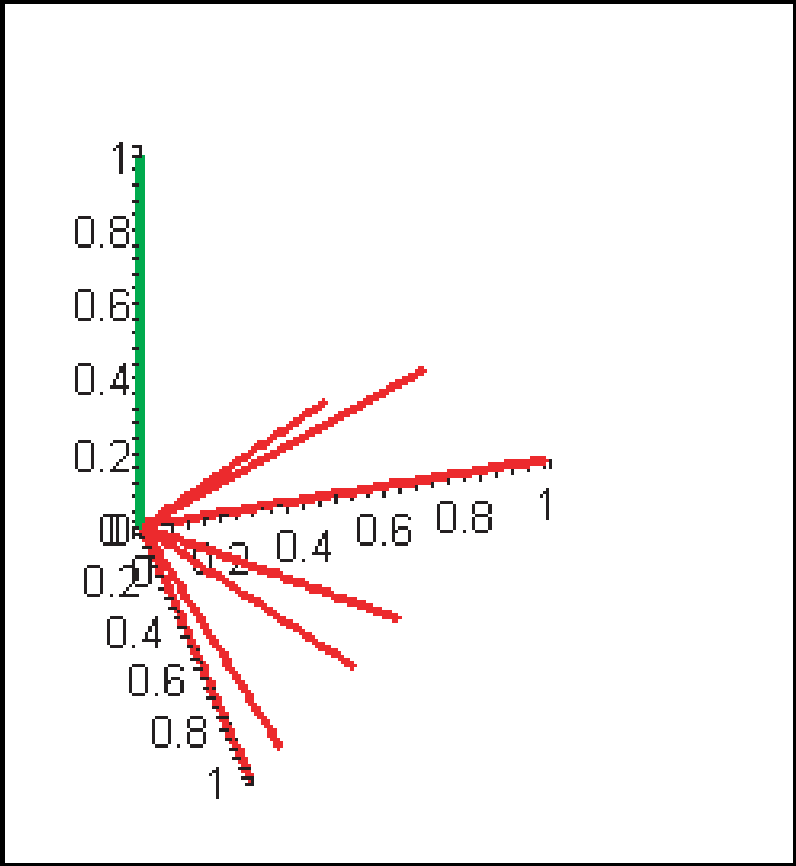
T1: Bab(y,ies,y's)
T2: Child(ren's)
T3: Guide
T4: Health
T5: Home
T6: Infant
T7: Proofing
T8: Safety
T9: Toddler

Documents

D1: Infant & Toddler First Aid
D2: Babies & Children's Room (For Your Home)
D3: Child Safety at Home
D4: Your Baby's Health & Safety : From Infant to Toddler
D5: Baby Proofing Basics
D6: Your Guide to Easy Rust Proofing
D7: Beanie Babies Collector's Guide

$$\mathbf{A} = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \quad \mathbf{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \end{bmatrix} = \begin{bmatrix} 0 \\ .5774 \\ 0 \\ .8944 \\ .7071 \\ 0 \\ .7071 \end{bmatrix}$$

Geometry of VSM for 3d vectors



vsmanimation.html

Latent Semantic Indexing (1990s)

Susan Dumais's improvement to VSM = LSI

Idea: use low-rank approximation to **A** to filter out noise

- Great Idea! 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.

  (Resource: USPTO `http://patft.uspto.gov/netahtml/srchnum.htm`)

Singular Value Decomposition

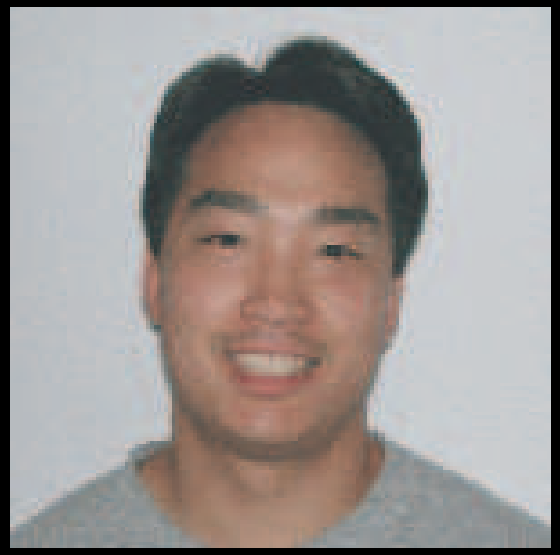
$\mathbf{A}_{m \times n}$: rank r term-by-document matrix

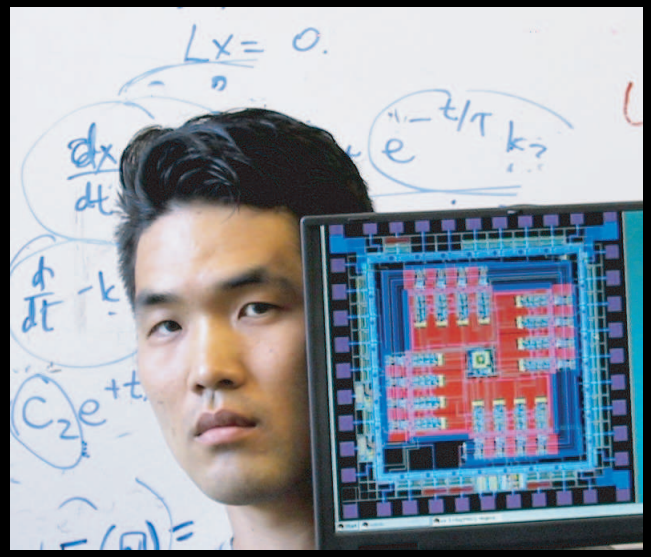
- SVD: $\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$
- LSI: use $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ in place of $\mathbf{A}$
- Why?
  - reduce storage when $k << r$
  - filter out uncertainty, so that performance on text mining tasks (e.g., query processing and clustering) improves

LSI Demos

- Telcordia LSI Demo: http://lsi.research.telcordia.com/lsi-bin/lsiQuery
- Netlib LSI Demo: http://www.netlib.org/cgi-bin/lsiBook

Nonnegative Matrix Factorization (2000)





Daniel Lee and Sebastian Seung's Nonnegative Matrix Factorization

Idea: use low-rank approximation with nonnegative factors to improve LSI

$$\underset{nonneg}{\mathbf{A}_k} = \underset{mixed}{\mathbf{U}_k} \; \underset{nonneg}{\Sigma_k} \; \underset{mixed}{\mathbf{V}_k^T}$$

$$\underset{nonneg}{\mathbf{A}_k} = \underset{nonneg}{\mathbf{W}_k} \; \underset{nonneg}{\mathbf{H}_k}$$

Properties of NMF

- can restrict **W**, **H** to be sparse
- $\mathbf{W}_k$, $\mathbf{H}_k \geq 0 \Rightarrow$ immediate interpretation (additive parts-based rep.)

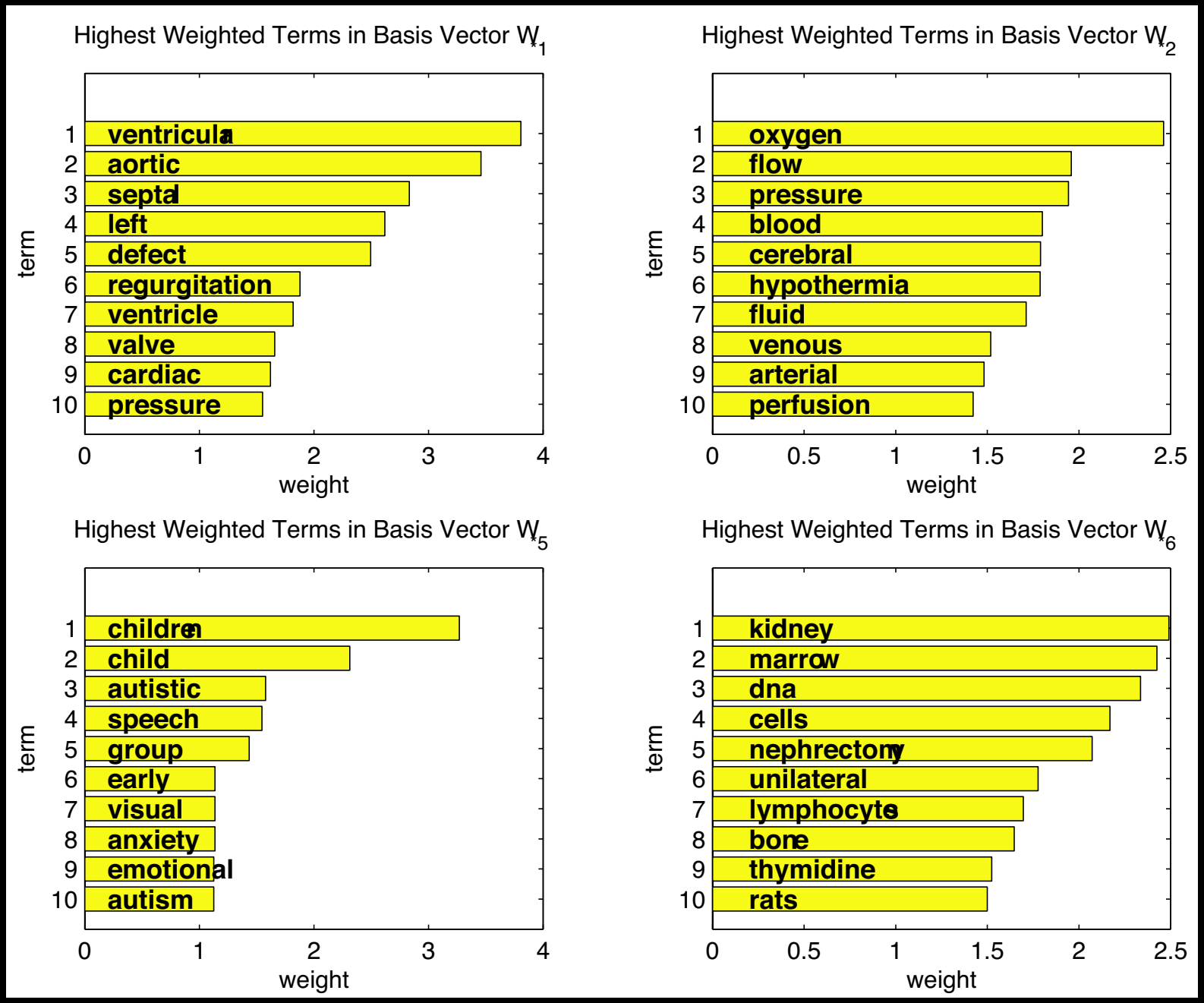
  EX: large w_{ij}'s $\Rightarrow$ basis vector $\mathbf{w}_i$ is mostly about terms j

  EX: h_{i1} how much doc_1 is pointing in the "direction" of topic vector $\mathbf{w}_i$

$$\mathbf{A}_k\mathbf{e}_1 = \mathbf{W}_k\mathbf{H}_{*1} = \begin{bmatrix} \vdots \\ \mathbf{w}_1 \\ \vdots \end{bmatrix} h_{11} + \begin{bmatrix} \vdots \\ \mathbf{w}_2 \\ \vdots \end{bmatrix} h_{21} + \cdots + \begin{bmatrix} \vdots \\ \mathbf{w}_k \\ \vdots \end{bmatrix} h_{k1}$$

Interpretation of Basis Vectors

MED dataset ($k = 10$)



Interpretation of Basis Vectors

MED dataset ($k = 10$)

$$\mathbf{doc}_5 \approx \begin{array}{c} \mathbf{w}_9 \\ \begin{pmatrix} \text{fatty} \\ \text{glucose} \\ \text{acids} \\ \text{ffa} \\ \text{insulin} \\ \vdots \end{pmatrix} \end{array} .1646 + \begin{array}{c} \mathbf{w}_6 \\ \begin{pmatrix} \text{kidney} \\ \text{marrow} \\ \text{dna} \\ \text{cells} \\ \text{nephr.} \\ \vdots \end{pmatrix} \end{array} .0103 + \begin{array}{c} \mathbf{w}_7 \\ \begin{pmatrix} \text{hormone} \\ \text{growth} \\ \text{hgh} \\ \text{pituitary} \\ \text{mg} \\ \vdots \end{pmatrix} \end{array} .0045 + \cdots$$

Computation of NMF

(Lee and Seung 2000)

MEAN SQUARED ERROR OBJECTIVE FUNCTION

$$\min \|\mathbf{A} - \mathbf{W}\mathbf{H}\|^2 \quad s.t. \quad \mathbf{W}, \mathbf{H} \geq 0$$

```
W = abs(randn(m,k));
H = abs(randn(k,n));
for i = 1 : maxiter
    H = H .* (W^T A) ./ (W^T W H + 10^-9);
    W = W .* (A H^T) ./ (W H H^T + 10^-9);
end
```

Many parameters affect performance (k, obj. function, sparsity constraints, algorithm, etc.).

— NMF is not unique!

Google Job Opportunities

Home

About Google

We're Hiring!
Main
All Openings
Top 10 Reasons
Culture
Benefits
Inside View
Work/Life Balance

Find on this site:
Search

Looking for interesting work that matters to millions of people?

Google's mission:
Organize the world's information and make it universally accessible and useful.

To make this vision a reality, Google is looking for exceptional people who like to develop innovative new products, especially software engineers and tech-savvy product managers. Are you exceptional at what you do? Do you:

- Thrive on working in small teams to develop innovative products?
- Enjoy developing efficient new algorithms for processing tremendous amounts of data?
- Think it would be fun to write distributed systems that run on thousands of computers?
- Live to have the results of your work used and depended upon by millions of people every day?

If you're an outstanding software developer, computer scientist or product manager, read on and consider sending your resume and a brief cover letter to greatpeople@google.com.

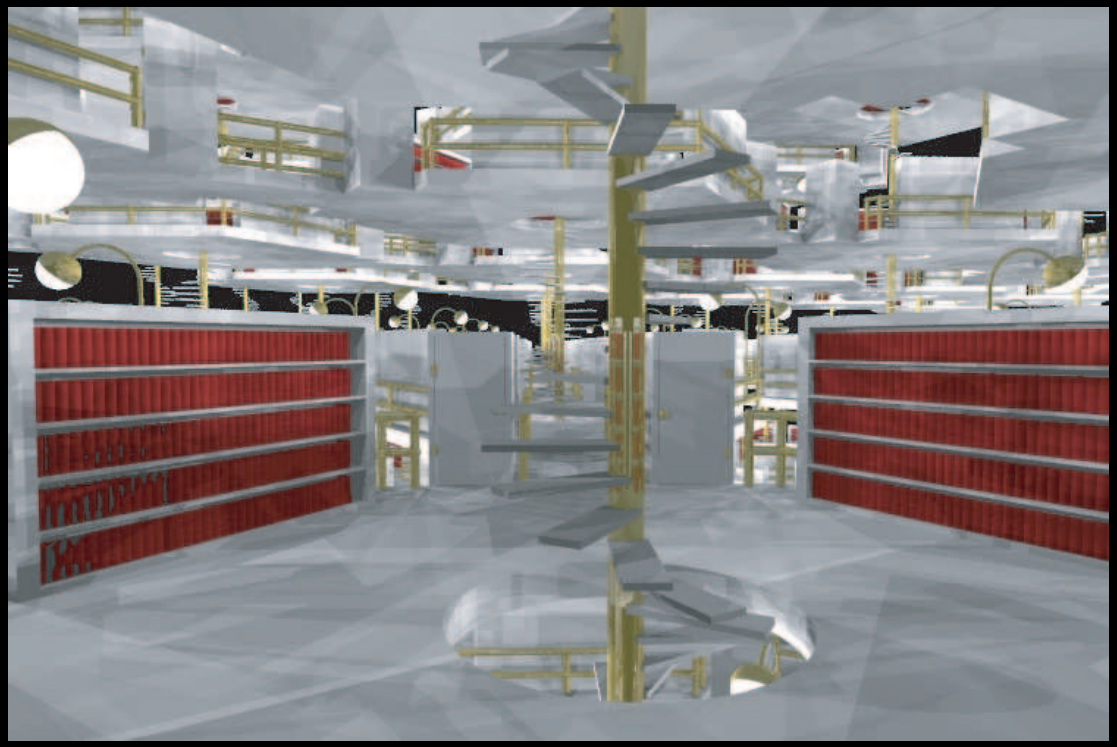
If you know others who fit in this category, help us improve Google by forwarding this message and URL to them. The URL for this page is:

http://www.google.com/jobs/great-people-needed.html

What is it like to work at Google?

Working at Google means solving fascinating problems and making a positive difference in tens of millions of lives every day. This work has opened up interesting new areas for us and presented challenges that are not only new to us, but new to everyone in computing. These new problems require exceptional thinking and technical expertise to solve, but their solutions could dramatically improve the accessibility of information for everyone in the world. Here's a sampling of the kinds of things we work on at Google:

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

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 - 550 billion pages

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 - 550 billion pages
- 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 - 550 billion pages
- 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 - 550 billion pages
- 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!

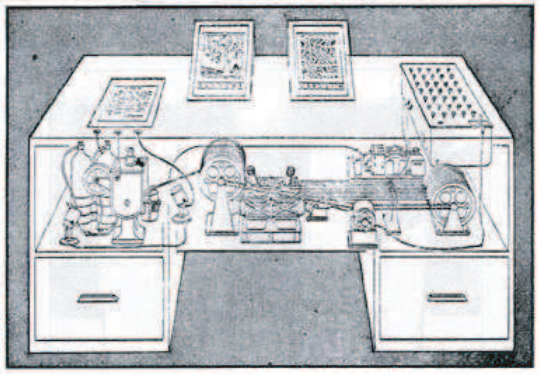
A Herculean Task!

Web Information Retrieval

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

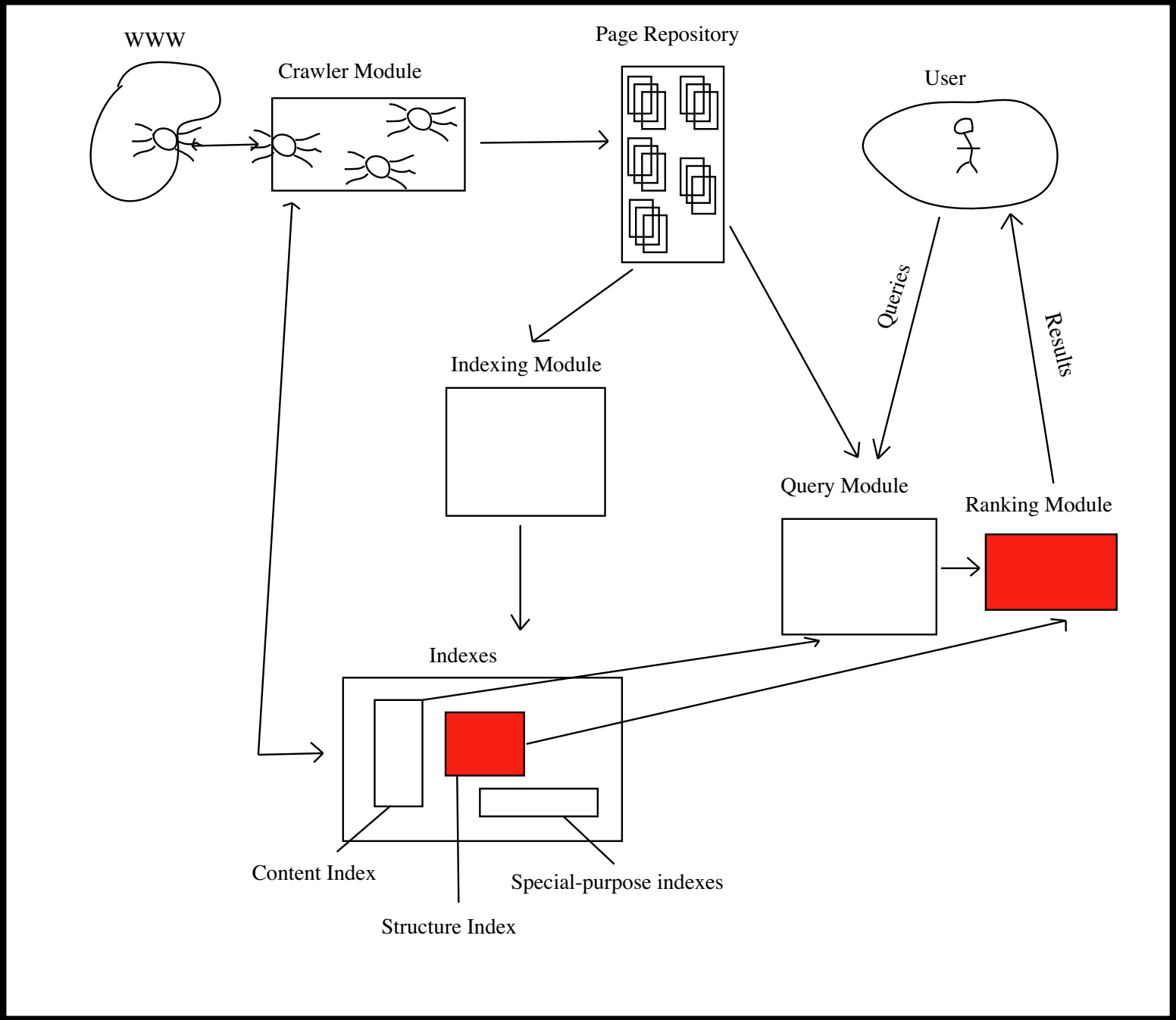
How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, each about 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 550 billion pages
- 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

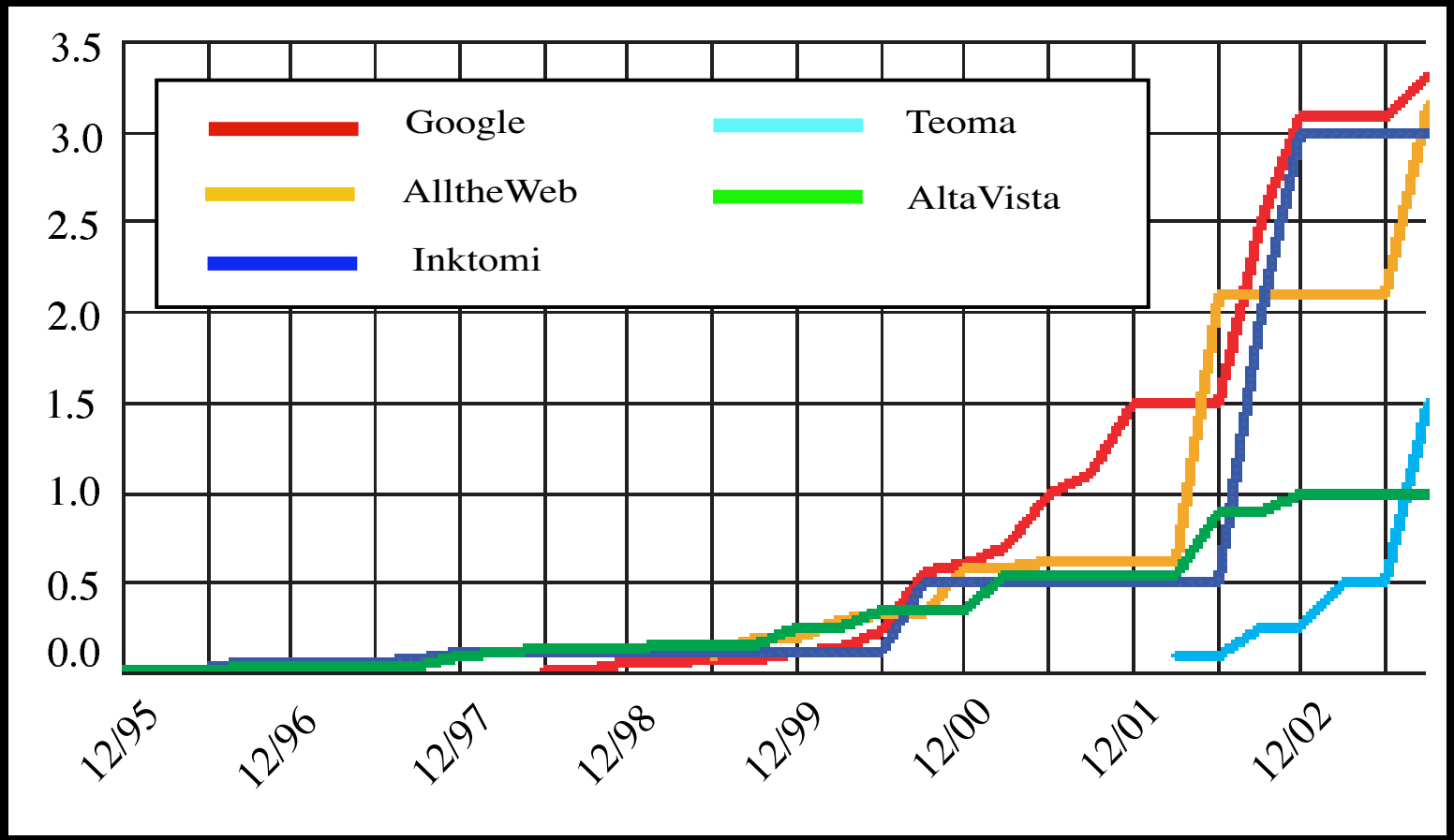


Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files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

Elements of a Web Search Engine



Indexing Wars



Actual Index King =

Internet Archive - http://web.archive.org

Search Stats—Google

- received over .5 billion searches per day in 2004
- stores an index of 8.1 billion webpages
- had over 60,000 servers in 2004
- estimated to use over 6,200 TB of disk space

Query Processing

Step 1: User enters query, i.e., aztec baby

Step 2: Inverted file consulted

- term 1 (aardvark) - 3, 117, 3961

  ⋮

- term 10 (aztec) - 3, 15, 19, 101, 673, 1199
- term 11 (baby) - 3, 31, 56, 94, 673, 909, 11114, 253791

  ⋮

- term m (zymurgy) - 1159223

Step 3: Relevant set identified, i.e. $(3, 673)$

Simple traditional engines stop here.

Modification to Inverted File

- add more features to inverted file by appending vector to each page identifier, i.e., [in title?, in descrip.?, # of occurrences]
- Modified inverted file

  - term 1 (aardvark) - 3 [0,0,3], 117 [1,1,10], 3961 [0,1,4]

    ⋮

  - term 10 (aztec) - 3 [1, 1, 27], 15 [0,0,1], 19 [1,1,21], 101 [0,1,7], 673 [0, 0, 3], 1199 [0,0,3]
  - term 11 (baby) - 3 [1, 1, 10], 31 [0,0,2], 56 [0,1,3], 94 [1,1,11], 673 [1, 1, 14], 909 [0,0,2], 11114 [1,1,22], 253791 [0,1,6]

    ⋮

  - term m (zymurgy) - 1159223 [1,1,9]

- IR score computed for each page in relevant set.

  EX: IR score (page 3) = $(1 + 1 + 27) \times (1 + 1 + 10) = 348$

  IR score (page 673) = $(0 + 0 + 3) \times (1 + 1 + 14) = 48$

  Early web engines stop here.

Problem = Ranking by IR score is not good enough.

CSC issues in Crawling and Indexing

- create parallel crawlers but avoid overlap
- ethical spidering
- how often to crawl pages, which pages to update
- best way to store huge inverted file
- how to efficiently update inverted file
- store the files across processors
- provide for parallel access
- create robust, failure-resistant system

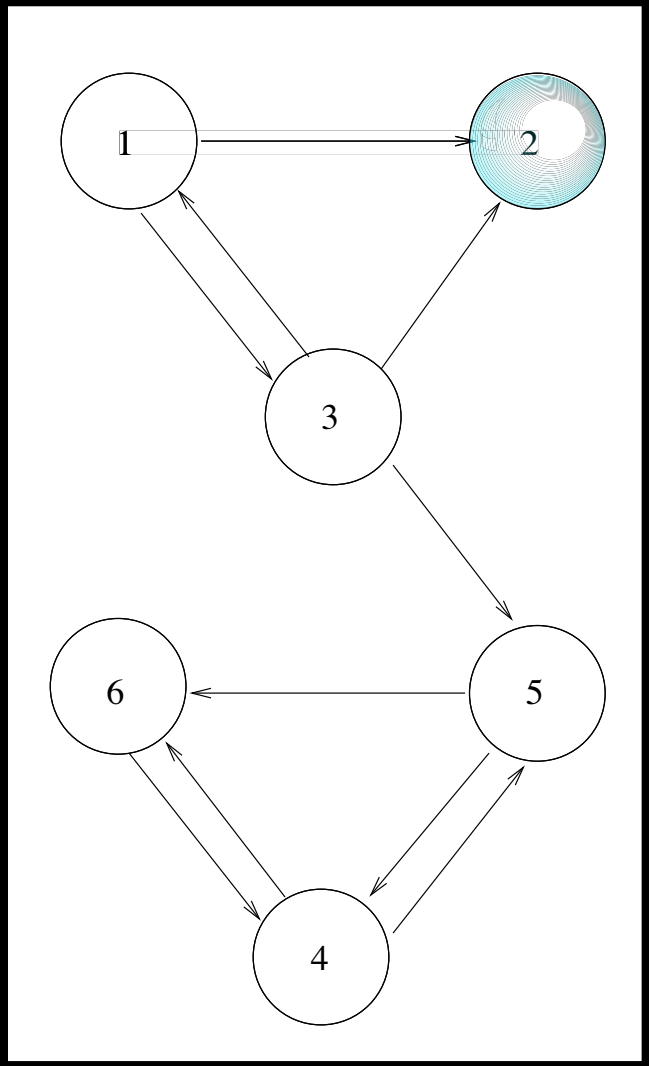
Link Analysis

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

PageRank - Brin and Page ⇒ Google

HITS - Kleinberg ⇒ TEOMA

The Web as a Graph



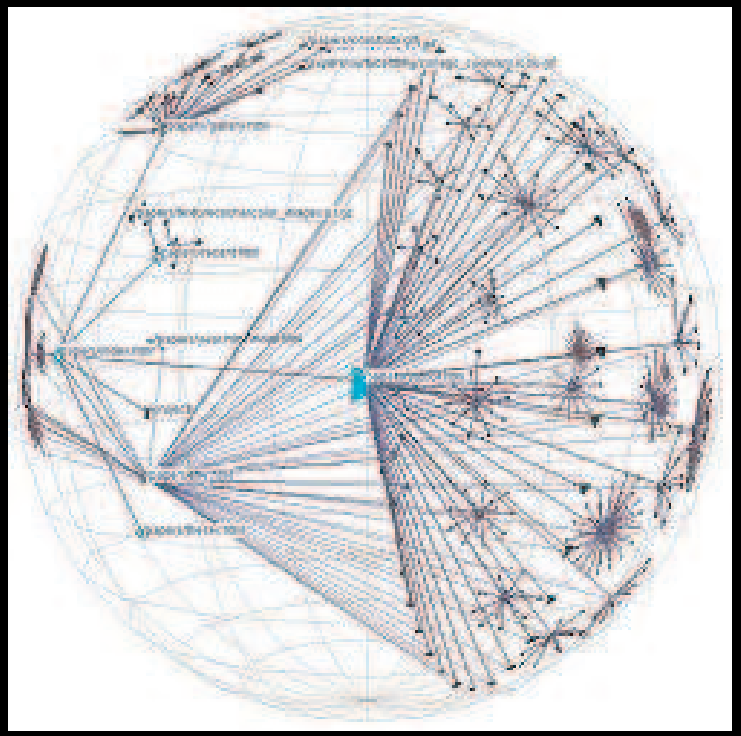
Nodes = webpages

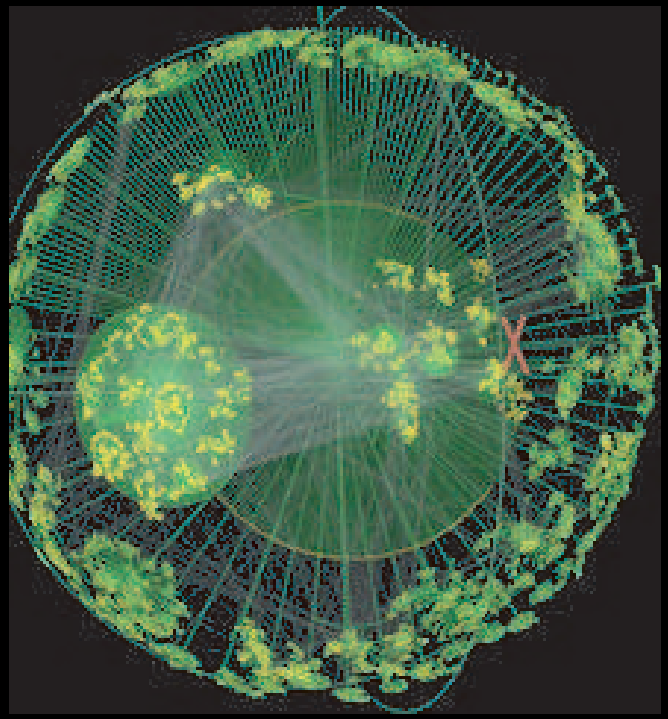
Arcs = hyperlinks

Web Graphs

CSC and MATH problems here:

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





How to Use Web Graph for Search

$\underline{Hyperlink}$ = **Recommendation**

- page with 20 recommmendations (inlinks) must be more important than page with 2 inlinks.
- but status of recommender matters.
  EX: letters of recommendation: 1 letter from Trump vs. 20 from unknown people
- but what if recommender is generous with recommendations?
  EX: suppose Trump has written over 40,000 letters.
- each inlink should be weighted to account for status of recommender and # of outlinks from that recommender

How to Use Web Graph for Search

$\underline{Hyperlink}$ = **Recommendation**

- page with 20 recommmendations (inlinks) must be more important than page with 2 inlinks.
- but status of recommender matters.
  EX: letters of recommendation: 1 letter from Trump vs. 20 from unknown people
- but what if recommender is generous with recommendations?
  EX: suppose Trump has written over 40,000 letters.
- each inlink should be weighted to account for status of recommender and # of outlinks from that recommender

PAGERANK - importance/popularity score given to each page

Home

All About Google

Help Central

Google Features

Services & Tools

Our Technology
- **Why Use Google**
- Benefits of Google

Find on this site:

Search

Google searches more sites more quickly, delivering the most relevant results.

Introduction

Google runs on a unique combination of advanced hardware and software. The speed you experience can be attributed in part to the efficiency of our search algorithm and partly to the thousands of low cost PC's we've networked together to create a superfast search engine.

The heart of our software is PageRank™, a system for ranking web pages developed by our founders Larry Page and Sergey Brin at Stanford University. And while we have dozens of engineers working to improve every aspect of Google on a daily basis, PageRank continues to provide the basis for all of our web search tools.

PageRank Explained

PageRank relies on the uniquely democratic nature of the web by using its

Ranking by PageRank

The PageRank Idea

(Sergey Brin & Lawrence Page 1998)

- Ranking is preassigned (An off-line calculation)
- Your page P has some rank $r(P)$
- Adjust $r(P)$ higher or lower depending on ranks of pages that point to P
- Importance is not just number, but $quality$ of in-links
  - role of outlinks relegated
  - much less sensitive to spamming

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
  - $r_1(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_0(P)}{|P|}$
    - $r_2(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_1(P)}{|P|}$

$$\ddots$$

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

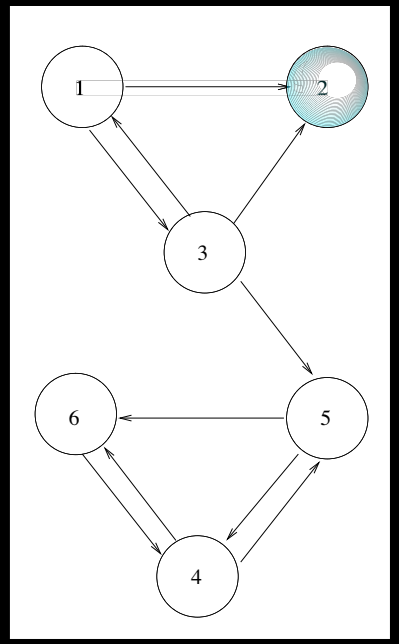
In Matrix Notation

After Step j

$$\boldsymbol{\pi}_j^T = \left[r_j(P_1),\, r_j(P_2),\, \cdots,\, r_j(P_n)\right]$$

$$\boldsymbol{\pi}_{j+1}^T = \boldsymbol{\pi}_j^T \mathbf{H} \quad \text{where} \quad h_{ij} = \begin{cases} \mathbf{1}/|P_i| & \text{if } i \to j \\ \mathbf{0} & \text{o.w.} \end{cases}$$

In Matrix Notation



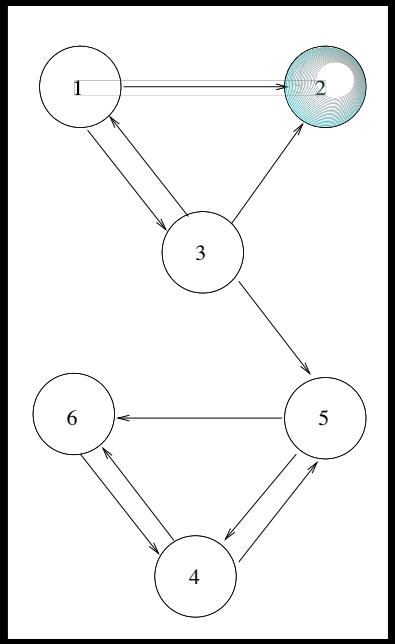
After Step j

$$\boldsymbol{\pi}_j^T = \left[r_j(P_1),\, r_j(P_2),\, \cdots,\, r_j(P_n)\right]$$

$$\boldsymbol{\pi}_{j+1}^T = \boldsymbol{\pi}_j^T \mathbf{H} \quad \text{where} \quad h_{ij} = \begin{cases} 1/|P_i| & \text{if } i \to j \\ 0 & \text{o.w.} \end{cases}$$

$$\mathbf{H} = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \end{array} \begin{array}{c} \begin{array}{cccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

In Matrix Notation



After Step j

$$\boldsymbol{\pi}_j^T = \left[r_j(P_1),\, r_j(P_2),\, \cdots,\, r_j(P_n)\right]$$

$$\boldsymbol{\pi}_{j+1}^T = \boldsymbol{\pi}_j^T \mathbf{H} \quad \text{where} \quad h_{ij} = \begin{cases} 1/|P_i| & \text{if } i \to j \\ 0 & \text{o.w.} \end{cases}$$

$$\mathbf{H} = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \end{array} \begin{array}{c} \begin{array}{cccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

PageRank = $\lim_{j \to \infty} \boldsymbol{\pi}_j^T = \boldsymbol{\pi}^T$ (provided limit exists)

It's Almost a Markov Chain

H has row sums = 1 for ND nodes, row sums = 0 for D nodes

In Matrix Notation

It's Almost a Markov Chain

- **H** has row sums = 1 for ND nodes, row sums = 0 for D nodes

In Matrix Notation

It's Almost a Markov Chain

- **H** has row sums = 1 for ND nodes, row sums = 0 for D nodes

Stochasticity Fix: **S** = **H** + **a**$\mathbf{v}^T$. (a_i=1 for $i \in D$, 0, o.w.)

In Matrix Notation

It's Almost a Markov Chain

- **H** has row sums = 1 for ND nodes, row sums = 0 for D nodes

Stochasticity Fix: **S** = **H** + **a**$\mathbf{v}^T$. (a_i=1 for $i \in D$, 0, o.w.)

$$\mathbf{S} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, where \ \mathbf{a} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{v}^T = 1/6 \ \mathbf{e}^T$$

In Matrix Notation

It's Almost a Markov Chain

- **H** has row sums = 1 for ND nodes, row sums = 0 for D nodes

Stochasticity Fix: **S** = **H** + **a**$\mathbf{v}^T$. (a_i=1 for $i \in D$, 0, o.w.)

$$\mathbf{S} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, where \ \mathbf{a} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{v}^T = 1/6 \ \mathbf{e}^T$$

- Each π_j^T is a probability distribution vector $\left(\sum_i r_j(P_i)=1\right)$
- $\pi_{j+1}^T = \pi_j^T \mathbf{S}$ is random walk on the graph defined by links
- $\pi^T = \lim_{j \to \infty} \pi_j^T$ = stationary probability distribution

Random Surfer

Web Surfer Randomly Clicks On Links (Back button not a link)

Long-run proportion of time on page P_i is π_i

Problems

Random Surfer

Web Surfer Randomly Clicks On Links (Back button not a link)

Long-run proportion of time on page P_i is π_i

Problems

Dead end page (nothing to click on) ($\boldsymbol{\pi}^T$ not well defined)

Could get trapped into a cycle ($P_i \rightarrow P_j \rightarrow P_i$) (No convergence)

Random Surfer

Web Surfer Randomly Clicks On Links (Back button not a link)

Long-run proportion of time on page P_i is π_i

Problems

Dead end page (nothing to click on) ($\boldsymbol{\pi}^T$ not well defined)

Could get trapped into a cycle ($P_i \rightarrow P_j \rightarrow P_i$) (No convergence)

Convergence

Markov chain must be irreducible and aperiodic

Random Surfer

Web Surfer Randomly Clicks On Links (Back button not a link)

Long-run proportion of time on page P_i is π_i

Problems

Dead end page (nothing to click on) ($\boldsymbol{\pi}^T$ not well defined)

Could get trapped into a cycle ($P_i \to P_j \to P_i$) (No convergence)

Convergence

Markov chain must be irreducible and aperiodic

DEFN: a chain is *irreducible* if every page is reachable from every other page.

DEFN: every *reducible* chain can be permuted to the form $\begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{bmatrix}$.

Random Surfer

Bored Surfer Enters Random URL

Irreducibility Fix: $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E}$ $\quad e_{ij} = 1/n \quad \alpha \approx .85$

$\mathbf{G} = \alpha\mathbf{H} + \alpha\,\mathbf{a}\,\mathbf{v}^T + (1-\alpha)\mathbf{E}$ (trivially irreducible)

- π^T is now guaranteed to exist and be unique and power method is guaranteed to converge to π^T.

Random Surfer

Bored Surfer Enters Random URL

Irreducibility Fix: $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E}$ $\quad e_{ij} = 1/n \quad \alpha \approx .85$

$\mathbf{G} = \alpha\mathbf{H} + \alpha\,\mathbf{a}\,\mathbf{v}^T + (1-\alpha)\mathbf{E}$ (trivially irreducible)

- π^T is now guaranteed to exist and be unique and power method is guaranteed to converge to π^T.

- Different $\mathbf{E} = \mathbf{e}\mathbf{v}^T$ and α allow customization & speedup, yet rank-one update maintained; $\mathbf{G} = \alpha\mathbf{H} + (\alpha\,\mathbf{a} + (1-\alpha)\,\mathbf{e})\mathbf{v}^T$

$$\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} = \begin{bmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{bmatrix}$$

Computing π^T

A Big Problem

Solve $\pi^T = \pi^T \mathbf{G}$ (stationary distribution vector)

$\pi^T(\mathbf{I} - \mathbf{G}) = \mathbf{0}$ (too big for direct solves)

CLEVE'S CORNER

THE WORLD'S LARGEST MATRIX COMPUTATION

Google's PageRank is an eigenvector of a matrix of order 2.7 billion.

BY CLEVE MOLER

One of the reasons why Google is such an effective search engine is the PageRank™ algorithm, developed by Google's founders, Larry Page and Sergey Brin, when they were graduate students at Stanford University. PageRank is determined entirely by the link structure of the Web. It is recomputed about once a month and does not involve any of the actual content of Web pages or of any individual query. Then, for any particular query, Google finds the pages on the Web that match that query and lists those pages in the order of their PageRank.

Imagine surfing the Web, going from page to page by randomly choosing an outgoing link from one page to get to the next. This can lead to dead ends at pages with no outgoing links, or cycles around cliques of interconnected pages. So, a certain fraction of the time, simply choose a random page from anywhere on the Web. This theoretical random walk of the Web is a *Markov chain* or *Markov process*. The limiting probability that a dedicated random surfer visits any particular page is its PageRank. A page has high rank if it has links to and from other pages with high rank.

Let W be the set of Web pages that can reached by following a chain of hyperlinks starting from a page at Google and let n be the number of pages in W. The set W actually varies with time, but in May 2002, n was about 2.7 billion. Let G be the n-by-n connectivity matrix of

It tells us that the largest eigenvalue of A is equal to one and that the corresponding eigenvector, which satisfies the equation

$$x = Ax,$$

exists and is unique to within a scaling factor. When this scaling factor is chosen so that

$$\sum_i x_i = 1$$

then x is the state vector of the Markov chain. The elements of x are Google's PageRank.

If the matrix were small enough to fit in MATLAB, one way to compute the eigenvector x would be to start with a good approximate solution, such as the PageRanks from the previous month, and simply repeat the assignment statement

```
x = Ax
```

until successive vectors agree to within specified tolerance. This is known as the power method and is about the only possible approach for very large n. I'm not sure how Google actually computes PageRank, but one step of the power method would require one pass over a database of Web pages, updating weighted reference counts generated by the hyperlinks between pages.

Computing π^T

A Big Problem

Solve $\pi^T = \pi^T \mathbf{G}$ (stationary distribution vector)

$\pi^T(\mathbf{I} - \mathbf{G}) = 0$ (too big for direct solves)

Start with $\pi_0^T = \mathbf{e}/n$ and iterate $\pi_{j+1}^T = \pi_j^T \mathbf{G}$ (power method)

Power Method to compute PageRank

$\boldsymbol{\pi}_0^T = \mathbf{e}^T/n$

until convergence, do

$$\boldsymbol{\pi}_{j+1}^T = \boldsymbol{\pi}_j^T \, \mathbf{G}$$

(dense computation)

end

Power Method to compute PageRank

$\boldsymbol{\pi}_0^T = \mathbf{e}^T/n$

until convergence, do

X $\boldsymbol{\pi}_{j+1}^T = \boldsymbol{\pi}_j^T\ \mathbf{G}$ (dense computation)

• $\boldsymbol{\pi}_{j+1}^T = \alpha\ \boldsymbol{\pi}_j^T\ \mathbf{S} + (1-\alpha)\ \boldsymbol{\pi}_j^T\ \mathbf{e}\ \mathbf{v}^T$ (sparser computation)

end

Power Method to compute PageRank

$\boldsymbol{\pi}_0^T = \mathbf{e}^T/n$

until convergence, do

X $\boldsymbol{\pi}_{j+1}^T = \boldsymbol{\pi}_j^T\ \mathbf{G}$ (dense computation)

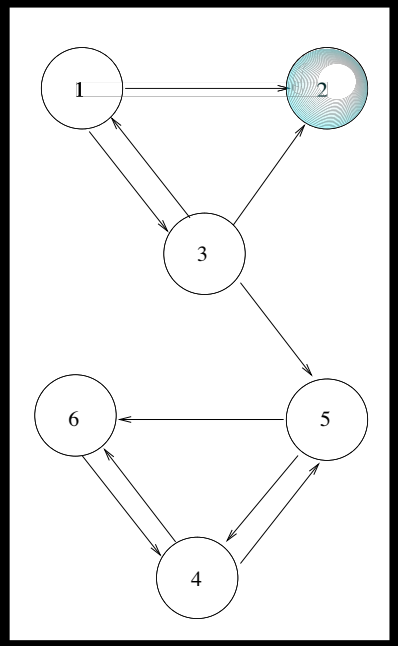
X $\boldsymbol{\pi}_{j+1}^T = \alpha\ \boldsymbol{\pi}_j^T\ \mathbf{S} + (1-\alpha)\ \boldsymbol{\pi}_j^T\ \mathbf{e}\ \mathbf{v}^T$ (sparser computation)

- $\boldsymbol{\pi}_{j+1}^T = \alpha\ \boldsymbol{\pi}_j^T\ \mathbf{H} + (\alpha\, \boldsymbol{\pi}_j^T\ \mathbf{a} + (1-\alpha))\ \mathbf{v}^T$ (even less computation)

end

- **H** is very, very sparse with about 3-10 nonzeros per row.
- $\Rightarrow$ one vector-matrix mult. is $O(nnz(\mathbf{H})) \approx O(n)$.

PageRank Example



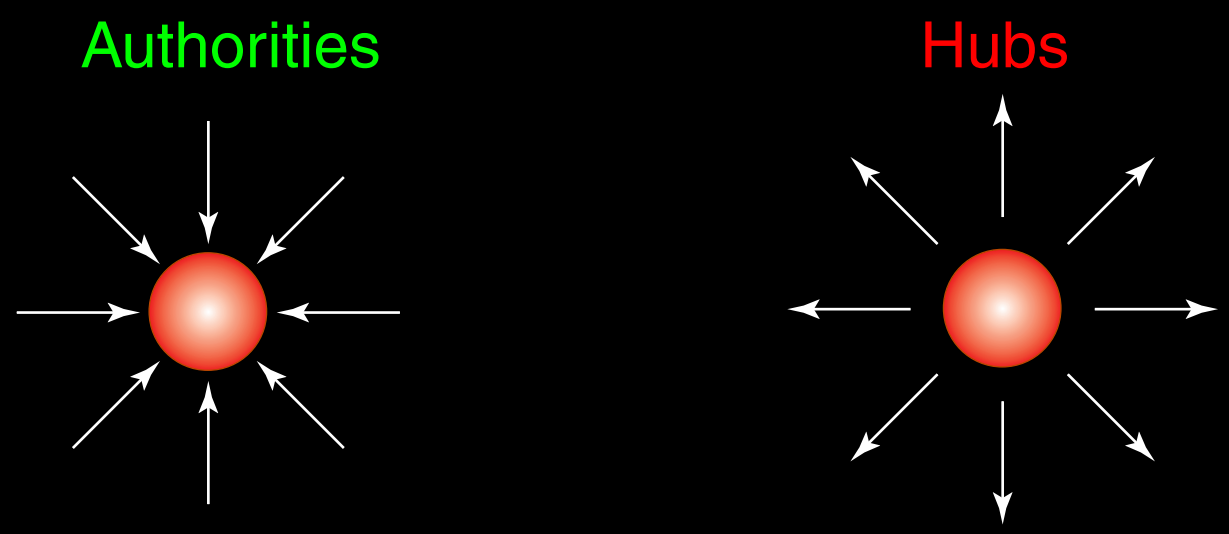
$$\pi^T = \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} .03721 & .05396 & .04151 & .3751 & .206 & .2862 \end{pmatrix} \end{array}$$

Global ranking of pages = $[\,4 \quad 6 \quad 5 \quad 2 \quad 3 \quad 1\,]$

Query-independent way of ranking relevant set

Ranking by HITS

- give each page 2 scores (hub and authority scores) instead of just 1.
- DEFN:



- pages can be both hubs and authorities (EX: ATL airport)
- Good hub pages point to good authority pages, and good authorities are pointed to by good hubs.

HITS - hub and authority score given to each page

HITS - (Hypertext Induced Topic Search)

ncaa basketball

Fin

Results
Relevant web pages

Showing 1-10 of about 3,255,000:

NCAA National Collegiate Athletic Association - Official Site
2004 **NCAA** Division I Men's **Basketball** Championship bracket announced...
www.ncaa.org/

Men's and Women's **Basketball** Polls
Division I Men's **Basketball** ... The **NCAA** does not conduct a poll for Division I men's **basketball** and the **NCAA**'s Division I Men's **Basketball** Committee...
www.ncaa.org/polls/m_w_basketball.html
[More Results from www.**ncaa**.org]

ESPN.com: Mens College **Basketball**
...to attend the EA SPORTS Maui Invitational **Basketball** Tournament, stay ... Wednesday ... 3:00 pm ... 1979 **NCAA** TOURNAMENT, MIDWEST REGIONAL 2ND...
sports.espn.go.com/ncb/index

Men's **Basketball** - **NCAA** Sports.com
Live Game Video **NCAA** March Madness on Demand brings you LIVE video of the Men's **Basketball** tournament. Division I...
www.ncaasports.com/basketball/mens

NCAA Basketball
Live Game Video **NCAA** March Madness on Demand brings you LIVE video of the Men's **Basketball** tournament. Division I Men's **Basketball**...
www.ncaasports.com/
[More Results from www.ncaasports.com]

D3hoops.com: The definitive resource for Division III men's and
The definitive resource for Division III men's and women's **basketball** ... previews: M | W Final Four: M | W Stats (**NCAA** site): M | W **NCAA** rankings: M...
www.d3hoops.com/

Women's **Basketball** Coaches Association
March 14 Selection Sunday for the **NCAA** Division I Women's **Basketball** Tournament March 16 NAIA DII Women's Championship March 19 **NCAA** DIII...
www.wbca.org/
[More Results from www.wbca.org]

CollegeRPI.com - College **Basketball** Rating Percentage Index (RPI)
The most accurate independent duplication of the **NCAA**'s Rating Percentage Index...
www.collegerpi.com/

College **Basketball** by CollegeHoopsnet.com
Player of the Week. **NCAA** Tournament. Conference Tourneys. **Basketball** Tickets. Recruiting Coverage. **Basketball** Store. NBA Draft...
www.collegehoopsnet.com/

CBS.SportsLine.com - **NCAA Basketball** Home
College **Basketball** coverage including **NCAA** news, scores, standings, stats, schedules, injuries, polls, team and player news, **NCAA basketball**...
www.sportsline.com/collegebasketball/

Refine
Suggestions to narrow your search

Basketball College

National Collegiate Athletic Association

Basketball Jersey In Ncaa

College Basketball News

Basketball Ncaa Rules

Basketball Rules

[Show All Refinements]

Resources
Link collections from experts and enthusiasts

College **Basketball** News: QuickSports.
sports.quickfound.net/...

SPL College **Basketball** Links & Tournament Contests
www.sportspl.com/...

Basketball News, NBA, **NCAA Basketball** - HeadlineSp...
www.headlinespot.com/...

Links for women's **basketball**
www.efn.org/...

Players CNNSI Player Rankings CNNSI Rosters Fox Sp...
www.sportgambler.com/...

NCAA BASKETBALL MEDIA LINKS
www.insidehoops.com/...

NCAA Basketball Links at Dharma Rose
www.darmarose.com/...

Girls Hoops, **Basketball** to look at, read about and...
www.girlshoops.org/...

MOP Squad Sports **NCAA Basketball**
www.mopsquad.com/...

NCAA Division 1 **basketball** Media on the Web
www.fastrackonline.net/...

Results Pages: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 >>

ncaa basketball

Fin

HITS Algorithm

Hypertext Induced Topic Search (J. Kleinberg 1998)

Determine Authority & Hub Scores

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

Successive Refinement

- Start with $h_i(0) = 1$ for all pages P_i
- Successively refine rankings
  - For $k = 1, 2, \ldots$

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

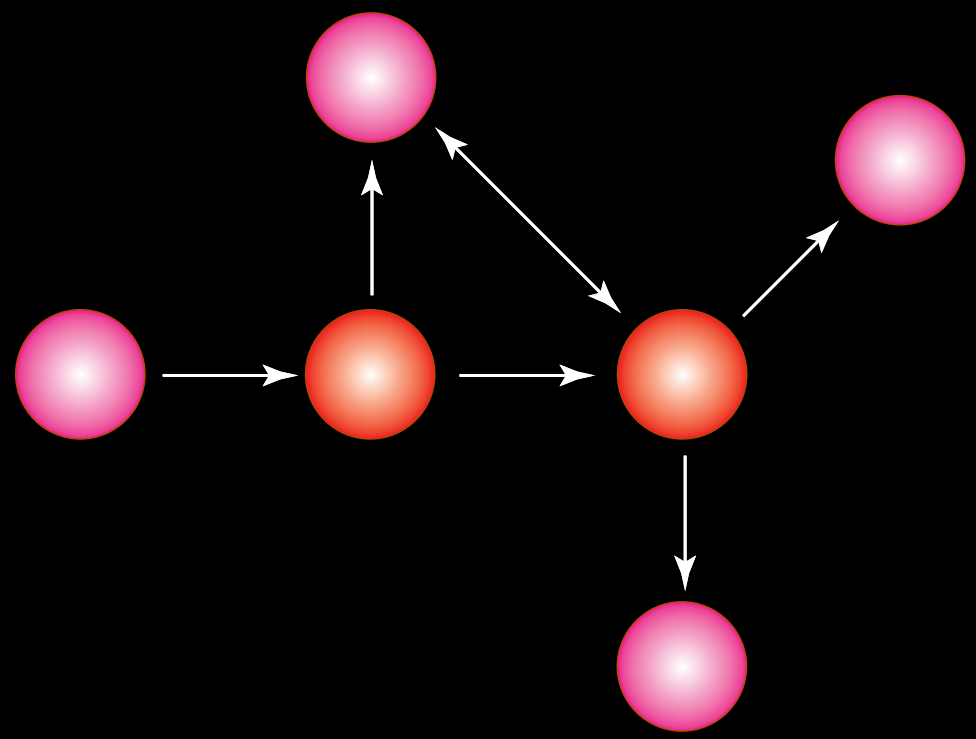
$$a_i(k) = \sum_{j:P_j \to P_i} h_j(k-1) \quad \Rightarrow \quad \mathbf{a}_k = \mathbf{L}^T \mathbf{h}_{k-1}$$

$$h_i(k) = \sum_{j:P_i \to P_j} a_j(k) \quad \Rightarrow \quad \mathbf{h}_k = \mathbf{L}\mathbf{a}_k$$

- $\mathbf{A} = \mathbf{L}^T\mathbf{L}$ $\quad \mathbf{a}_k = \mathbf{A}\mathbf{a}_{k-1} \to$ e-vector
- $\mathbf{H} = \mathbf{L}\mathbf{L}^T$ $\quad \mathbf{h}_k = \mathbf{H}\mathbf{h}_{k-1} \to$ e-vector

HITS Neighborhood Graph

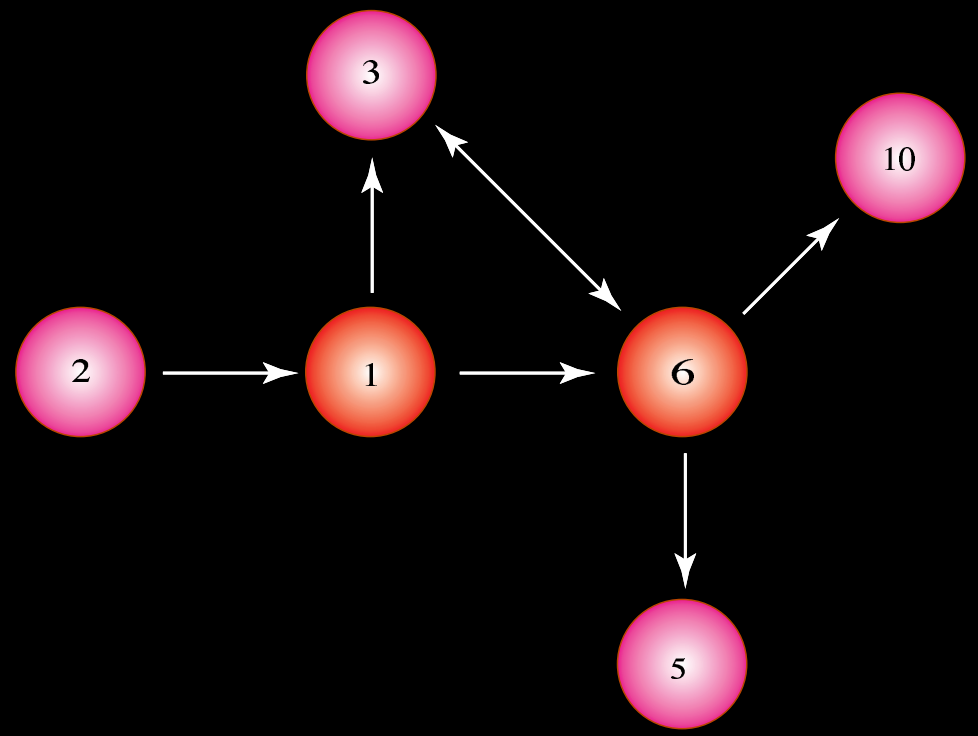
1. Find relevant set by consulting inverted file
2. Build neighborhood graph



3. Compute authority & hub scores for just the neighborhood

HITS Example

1. Relevant set = $[1, 6]$
2. Neighborhood graph N



3. Compute authority & hub scores.

Adjacency matrix for N = **L** =

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

HITS Example (cont.)

Authority matrix **A** = $\mathbf{L}^T\mathbf{L}$

Hub matrix **H** = $\mathbf{L}\mathbf{L}^T$

$$\mathbf{L}^T\mathbf{L} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}, \mathbf{L}\mathbf{L}^T = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Authority score vector **a**

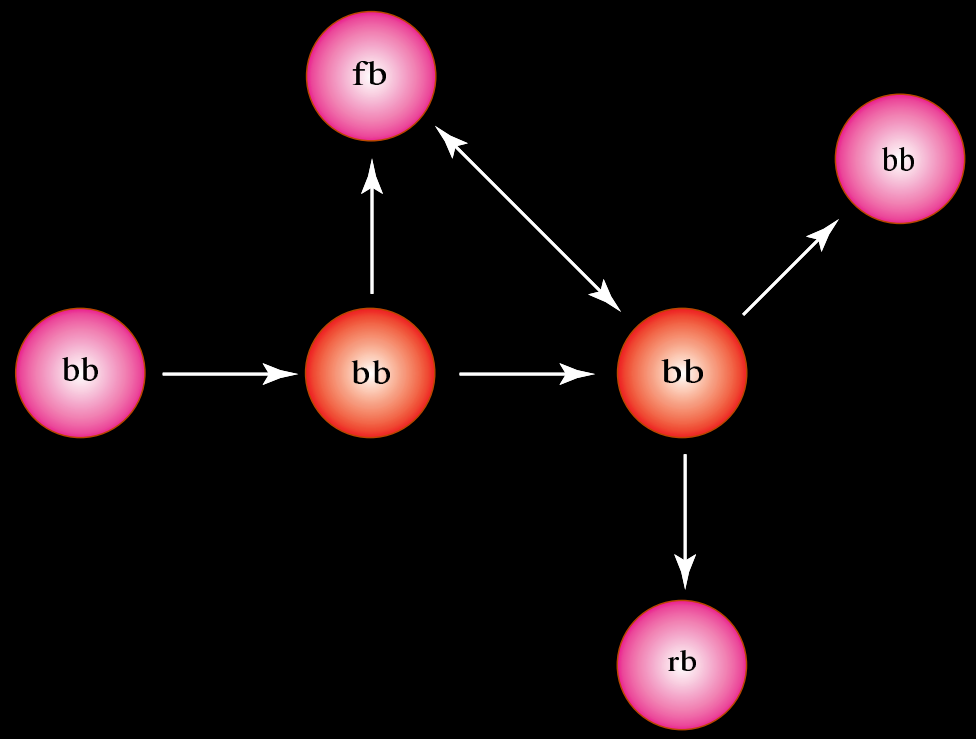
$$\mathbf{a}^T = \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & .3660 & .1340 & .5 & 0 \end{pmatrix} \end{array}$$

Hub score vector **h**

$$\mathbf{h}^T = \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} .3660 & 0 & .2113 & 0 & .2113 & .2113 \end{pmatrix} \end{array}$$

CSC and MATH Issues with HITS

- how to form N and fix topic drift problem



- incorporating weights into **L** matrix
- fast eigenvector computation, beating the power method
- updating **L**, **h**, and **a** for query-independent HITS

Power of Word of Mouth

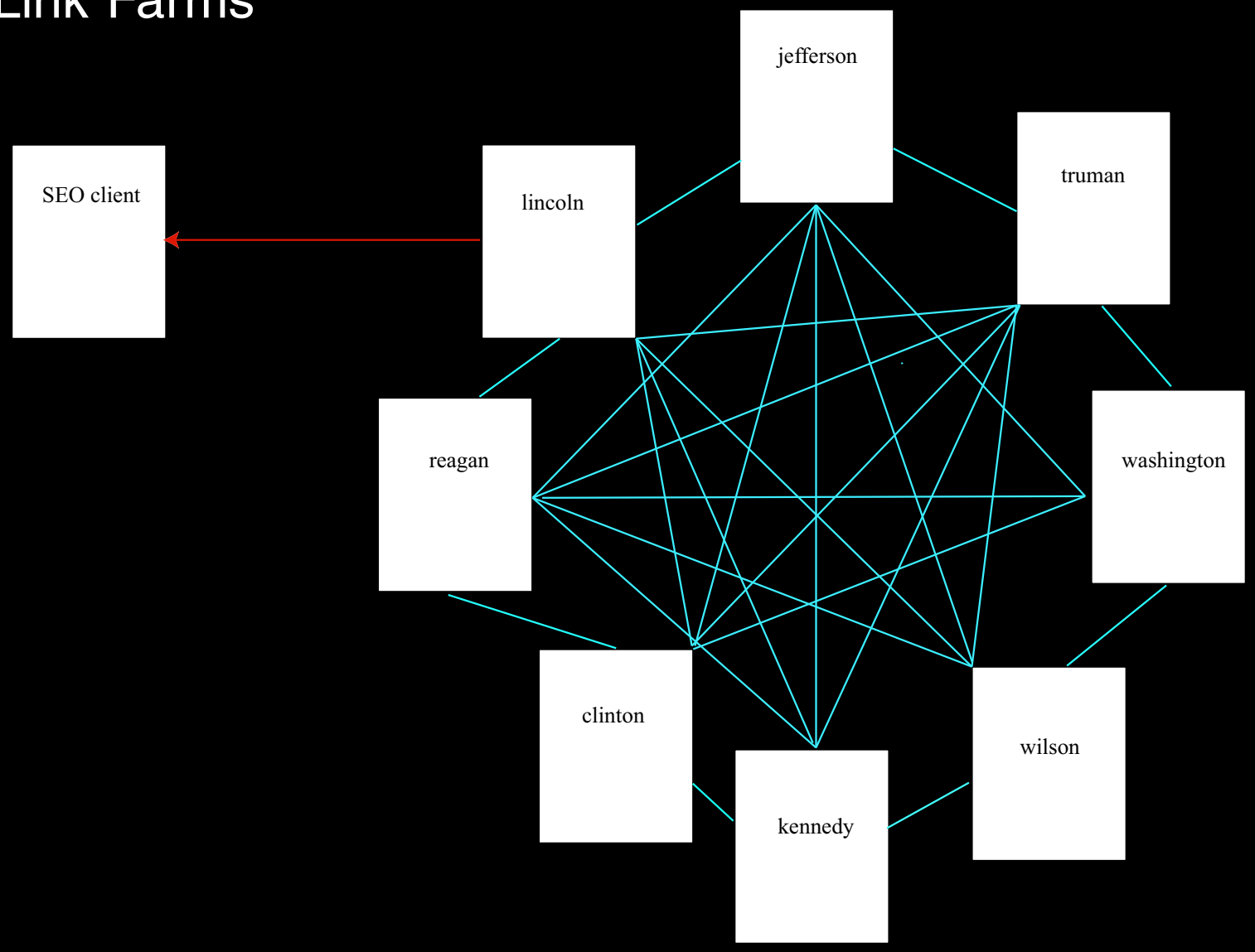
Other Rankings

- Consensus Ranking
- Traffic Ranking: www.alexa.com

Web Search Problems

Spam

- 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 rumors of a peaceful solution to

World-Wide

■ **BUSH IS PREPARING to present Congress a huge bill for Iraq costs.** The total could run to $95 billion depending on the length of the possible war and occupation. As horse-trading began at the U.N. to win support for a war resolution, the president again made clear he intends to act with or without the world body's imprimatur. Arms inspectors said Baghdad provided new data, including a report of a possible biological bomb. Gen. Franks assumed command of the war-operations center in Qatar. Allied warplanes are aggressively taking out missile sites that could threaten the allied troop build-up. (Column 4 and Pages A4 and A6)

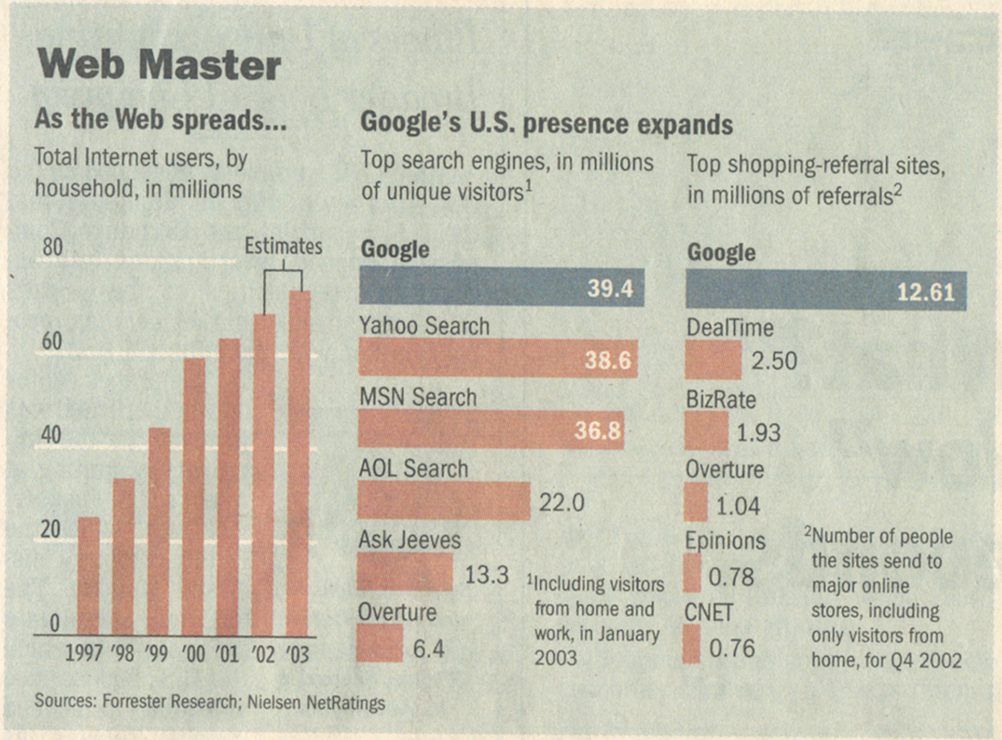
Turkey's parliament debated legislation to let the U.S. deploy 62,000 to open a northern front. Kurdish soldiers lined roads in a show of force as U.S. officials traveled into Iraq's north for an opposition conference.

* * *

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

* * *

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



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

By GREG JAFFE
And JOHN D. MCKINNON

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

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

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

Cat and Mouse

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

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

Exoticleatherwear Gets Cut Off

By MICHAEL TOTTY
And MYLENE MANGALINDAN

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

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

The techniques worked for a

Web Sites Fight for Prime Real Estate on Google

Continued From First Page

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

Google won't comment on the case. In court filings, the company said SearchKing "engaged in behavior that would lower the quality of Google search results" and alter the company's ranking system.

Google, a closely held company founded by Stanford University graduate students Sergey Brin and Larry Page, says Web companies that want to rank high should concentrate on improving their Web pages rather than gaming its system. "When people try to take scoring into their own hands, that turns into a worse experience for users," says Matt Cutts, a Google software engineer.

Coding Trickery

Efforts to outfox the search engines have been around since search engines first became popular in the early 1990s. Early tricks included stuffing thousands of widely used search terms in hidden coding, called "metatags." The coding fools a search engine into identifying a site with popular words and phrases that may not actually appear on the site.

Another gimmick was hiding words or terms against a same-color background. The hidden coding deceived search engines that relied heavily on the number of times a word or phrase appeared in ranking a site. But Google's system, based on links, wasn't fooled.

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

That didn't stop search optimizers from finding other ways to outfox the system. Attempts to manipulate Google's results even became a sport, called Google-bombing. Prankster would try to creating Web sites that were nothing more than collections of links to the clients' site, called "link farms." Since Google ranks a site largely by how many links or "votes" it gets, the link farms could boost a site's popularity.

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

Ms. Holman, the leatherwear retailer, discovered the consequences of trying to fool Google. The 42-year-old hospital laboratory technician, who learned computer skills by troubleshooting her hospital's equipment, operates her online apparel store as a side business that she hopes can someday replace her day job.

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

When she launched her Exotic Leather Wear store from her home in Mesa, Ariz., she quickly learned the importance of appearing near the top of search-engine results, especially on Google. She boned up on search techniques, visiting online discussion groups dedicated to search engines and reading what material she could find on the Web.

At first, Ms. Holman limited herself to modest changes, such as loading her page with hidden metatag coding that would help steer a search toward her site when a user entered words such as "haltertops" or "leather miniskirts." Since Google doesn't give much weight to metatags in determining its rankings, the efforts had little effect on her search results.

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

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

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

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

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

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

Crackdown

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

To stay one step ahead of the Web

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

Bob Massa, SearchKing's chief executive, last August launched the PR Ad Network as a way to capitalize on Google's page-ranking system, known as PageRank. PageRank rates Web sites on a scale of one to 10 based on their popularity, and the rankings can be viewed by Web users if they install special Google software. PR Ad Network sells ads that are priced according to a site's PageRank, with higher-ranked sites commanding higher prices. When a site buys an advertising link on a highly ranked site, the ad buyer could see its ratings improve because of the greater weight Google gives to that link.

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

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

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

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

Home Depot E
Amid First Qu

By Chad Terhune

ATLANTA—**Home Depot** Inc. repo
fiscal fourth-quarter earnings decli
3.4% on disappointing sales.

Speaking to investors and indu
analysts, the company's chairman
chief executive, Bob Nardelli,
Home Depot is prepared to win b
dissatisfied customers and answe
competitive challenge from its chie
val with remodeled stores, increased
ventory and improved customer serv

The nation's largest home-impr
ment retailer said net income for the q
ter ended Feb. 2 decreased to $686 mil
or 30 cents a share, from $710 million
30 cents a share, a year earlier. Sales
2% to $13.21 billion from $13.49 billion,
first quarterly sales decline in the co
ny's 24-year history. Home Depot n
the latest quarter was a week shorter t
a year earlier. Using comparable 13-v
periods, the company said quarterly s
increased 5% and net income rose 8.

Same-store sales, or sales at st
open at least a year, declined 6% in
quarter. Home Depot said stronger s
last month offset a disastrous Decem
and helped the retailer avoid its ea
estimate that same-store sales could
as much as 10%. In 4 p.m. New
Stock Exchange composite trading, H
Depot shares rose 66 cents to $22.84

Fiat Patriarch
Is Set to Becom

By Alessandra Galloni

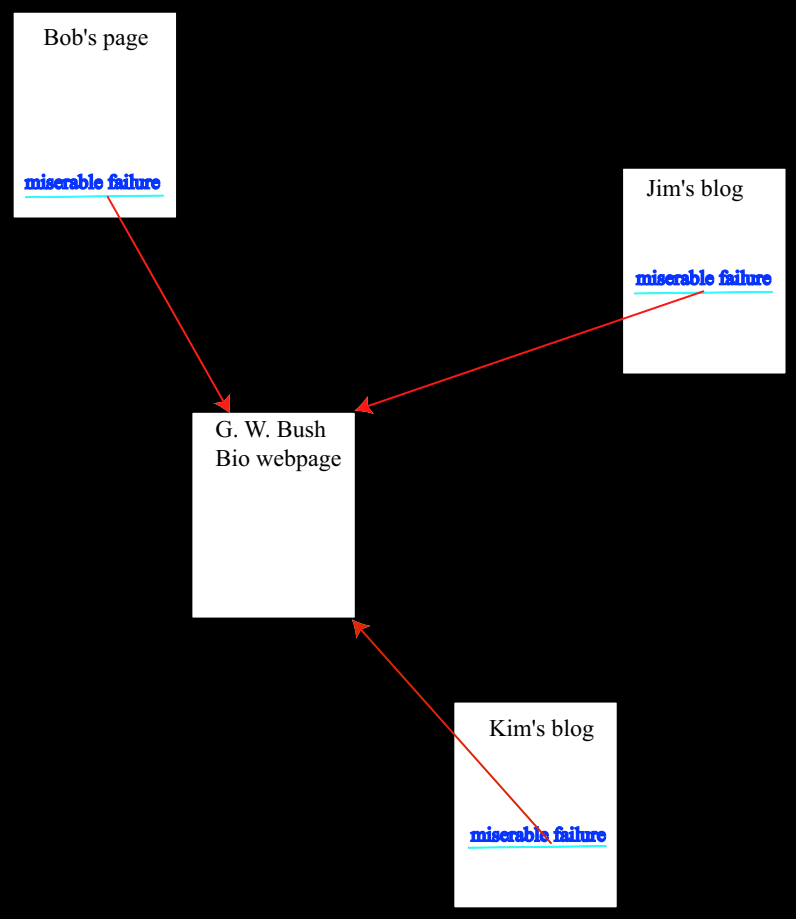
ROME—Umberto Agnelli is due t
named **Fiat** SpA chairman on Friday,
ping into the driver's seat as the Italian
glomerate works on an 11th-hour refin
ing of its unprofitable car unit.

Mr. Agnelli, the 68-year-old broth
Fiat patriarch Gianni Agnelli, who
last month, was widely expected to
over from current chairman, P
Fresco, later this year. But Mr. Fr
who has served as chairman since

Web Search Problems

Spam

- Link Farms
- Google Bombs



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.

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.

Bush has been the target of similar pranks before

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)".

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 this to a friend Printable version

LINKS TO MORE AMERICAS STORIES

Select Go

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

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.

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

Web Search Problems

Spam

- Link Farms
- Google Bombs
  - — search for algorithmic solutions that scale up

Social: getting surfers to use relevance feedback

- Specialized Tools: Froogle, Scholar

Web Search Problems

Spam

- Link Farms
- Google Bombs

Social: getting surfers to use relevance feedback

- Specialized Tools: Froogle, Scholar
- Personalization: www.a9.com

Web Search Problems

Spam

- Link Farms
- Google Bombs

Social: getting surfers to use relevance feedback

- Specialized Tools: Froogle, Scholar
- Personalization: www.a9.com

No human interaction, no librarian for mind-reading

Bad Results because ...

User's Fault

- poor query
- typo

Bad Results because ...

User's Fault

- poor query
- typo

Engine's Fault

- spam
- small index

Bad Results because ...

User's Fault

— poor query

— typo

Engine's Fault

— spam

— small index

Web Community's Fault

— no quality pages posted on user's query $\Rightarrow$ do your part

Innovation

- metatag for Library of Congress # (“nothing like a good book”)

Innovation

- metatag for Library of Congress # (“nothing like a good book”)
- connecting databases: maps.google.com, www.search.ch

Innovation

- metatag for Library of Congress # (“nothing like a good book”)
- connecting databases: maps.google.com, www.search.ch
- phonetic search in audio collections

Innovation

- metatag for Library of Congress # (“nothing like a good book”)
- connecting databases: maps.google.com, www.search.ch
- phonetic search in audio collections
- relevance feedback
  - — edit distance for typos
  - — synonyms (find similar terms using VSM)
  - — clustering (find similar docs): www.kartoo.com, www.vivisimo.com

Innovation

- metatag for Library of Congress # (“nothing like a good book”)
- connecting databases: maps.google.com, www.search.ch
- phonetic search in audio collections
- relevance feedback
  - — edit distance for typos
  - — synonyms (find similar terms using VSM)
  - — clustering (find similar docs): www.kartoo.com, www.vivisimo.com
- image search: google

Conclusions

- Link Analysis has drastically improved web search!
- There are many exciting open problems for computational scientists to solve.
- Often the challenge lies not in the modeling or theory, but in the massive scale of the problem.
- The continual battle between search engines and search engine optimizers means that methods must constantly adapt and innovate.
- There is huge financial potential for industrious entrepreneurs!