Overview of Markov-based Web IR systems

Amy Langville
Carl Meyer

Department of Mathematics
North Carolina State University
Raleigh, NC

Outline

- Background (25 %)
- HITS (20 %)
- PageRank (55 %)

World’s Largest Collection = WWW

Unique Features:

- Immense – some estimate > 10 billion pages
- Dynamic
- Growing – exponential, yet anticipated slowdown
- No editorial review process – broken links, redundancy, ...
- Business potential – leads to spamming
- Hyperlinks !

Some WWW Stats from Google

June 2002:

- Each page averages about 10K
- Currently indexes 3.08 billion webpages (July 2003)
- Receives 150 million search requests/day
- Serves 4000 searches/sec during peak times
- Uses 15,000 computers

Exploiting the Hyperlinks

Indexing and Ranking

- Must index key terms on each page
  - Robots crawl the web — software does indexing

- Inverted file structure
  - $Term_1 \longrightarrow P_i, P_j, \ldots$
  - $Term_2 \longrightarrow P_k, P_l, \ldots$

  $\vdots$

- Attach an importance rating to $P_i, P_j, P_k, P_l, \ldots$

- Direct query matching
  - $Q = Term_1, Term_2, \ldots$ produces $P_i, P_j, P_k, P_l, \ldots$

- Return $P_i, P_j, P_k, P_l, \ldots$ to user in order of importance

Search Engine Showdown
The Users' Guide
to Web Searching

Home Chart Reviews Statistics Learn Directories Search

Search Engine Statistics:
Database Total Size Estimates

by Greg R. Notess.

Search Engine	Showdown Estimate (millions)	Claim (millions)
Google	3,033	3,083
AlltheWeb	2,106	2,112
AltaVista	1,689	1,000
WiseNut	1,453	1,500
Hotbot	1,147	3,000
MSN Search	1,018	3,000
Teoma	1,015	500
NLResearch	733	125
Gigablast	275	150

Data from: Dec. 31, 2002

Based on AlltheWeb reported size and percentages from relative size showdown

AlltheWeb: 2,106,156,957 reported

The table above gives the Showdown Estimate and recent claims as to how many millions of Web pages have been indexed and included in the various search engines' databases. These estimates are based on exact counts obtained from AlltheWeb on the date of the comparison, and those numbers are

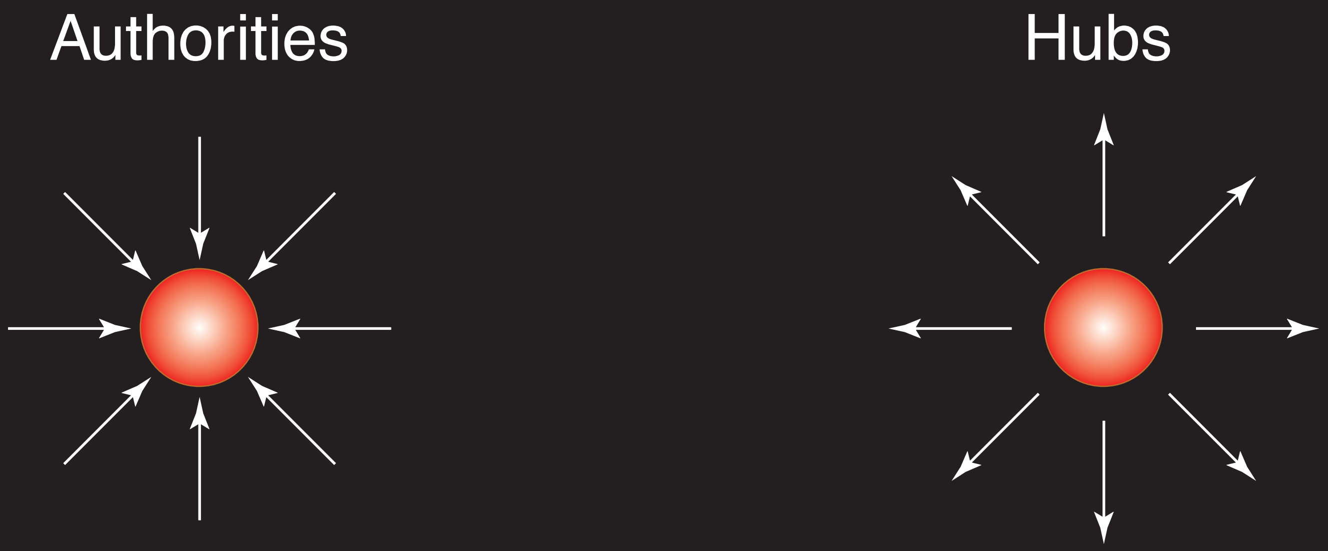
Focus of this Talk

How to attach importance to

rank

direct–matched pages ?

One Approach to Ranking: HITS



Authorities

Hubs

- Good hub pages point to good authority pages
- Good authorities are pointed to by good hubs

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

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

- Successively refine rankings
  - For $k = 1, 2, \ldots$

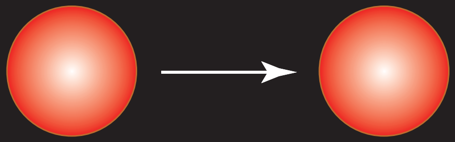
$$a_i(k) = \sum_{j:P_j \rightarrow P_i} h_j(k-1) \quad \Rightarrow \quad \mathbf{a}_k = \mathbf{L}^T \mathbf{h}_{k-1}$$

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

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

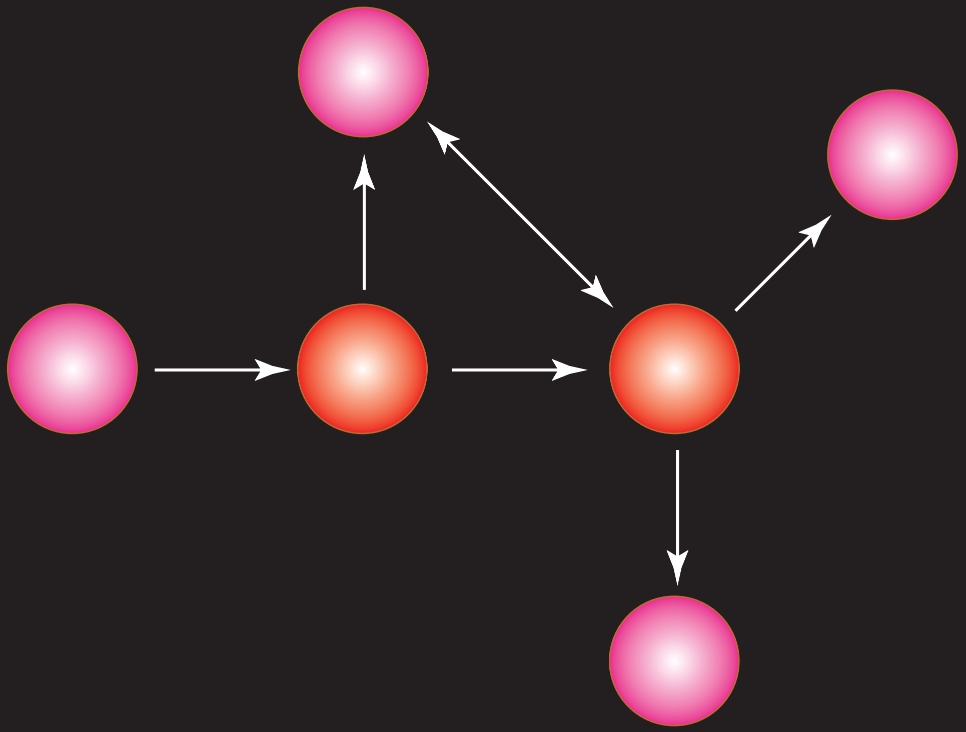
Compromise

1. Do direct query matching



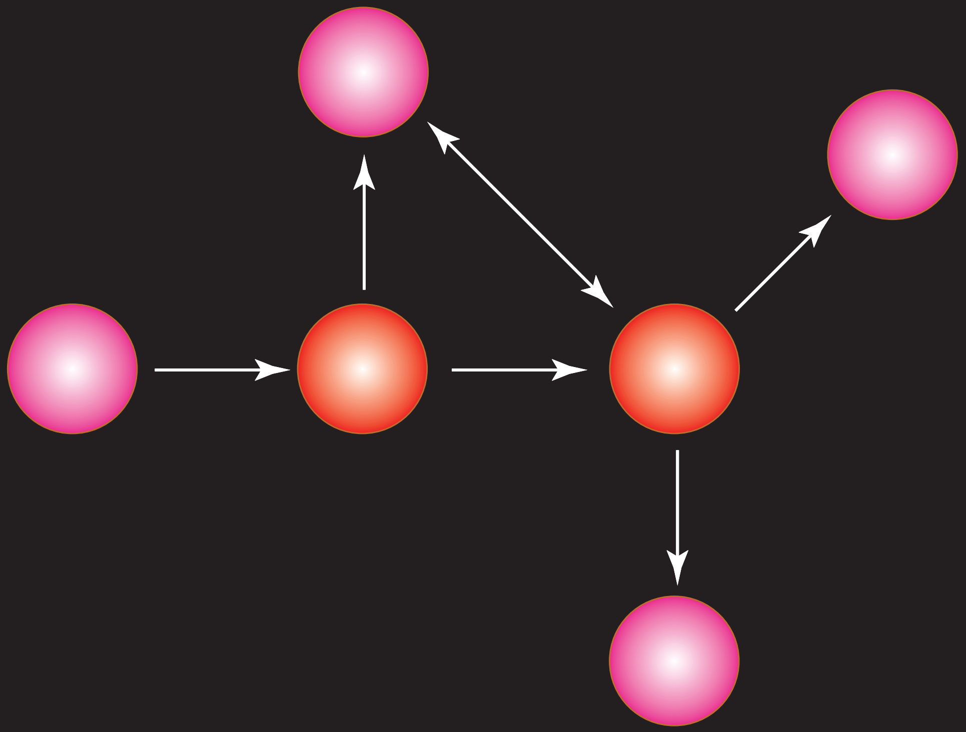
Compromise

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
- Some flexibility

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

Remnants of HITS in Citeseer

Approach to Ranking: 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

- $\pi_j^T = \left[r_j(P_1), r_j(P_2), \cdots, r_j(P_n)\right]$
- $\pi_{j+1}^T = \pi_j^T \mathbf{P}$ where $p_{ij} = \begin{cases} 1/|P_i| & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$
- PageRank $= \lim_{j\to\infty} \pi_j^T = \pi^T$ (provided limit exists)

It's A Markov Chain

- $\mathbf{P} = \left[p_{ij}\right]$ is a stochastic matrix (row sums = 1)
- Each π_j^T is a probability distribution vector $\left(\sum_i r_j(P_i) = 1\right)$
- $\pi_{j+1}^T = \pi_j^T \mathbf{P}$ 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

- Long-run proportion of time on page P_i is $\boldsymbol{\pi}_i$

(Back button is not a link. But, Fagin et.al. 2000 done back button modeling.)

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

Bored Surfer Enters Random URL

- Replace $\mathbf{P}$ by $\widetilde{\mathbf{P}} = \alpha\mathbf{P} + (1-\alpha)\mathbf{E}$ where $e_{ij} = 1/n$ $\alpha \approx .85$
  - Different $\mathbf{E}$'s and α's allow customization & speedup

PageRank becomes . . .

Computing π^T for irreducible, aperiodic Markov chain with transition probability matrix $\widetilde{\mathbf{P}} = \alpha\,\mathbf{P} + (1-\alpha)/n\,\mathbf{e}\,\mathbf{e}^T$.

- Eigenvector approach: Solve $\pi^T = \pi^T\widetilde{\mathbf{P}}$ (stationary vector)

  $\Rightarrow$ Power Method: $\pi_{j+1}^T = \pi_j^T\widetilde{\mathbf{P}} = \alpha\,\pi_j^T\mathbf{P} + (1-\alpha)/n\,\mathbf{e}^T$

- Linear system approach: $\pi^T(\mathbf{I} - \widetilde{\mathbf{P}}) = 0$

  — After some algebra: $\pi^T(\mathbf{I} - \alpha\,\mathbf{P}) = 1/n\,\mathbf{e}^T$ (too big for direct solves)

"The World's Largest Matrix Computation" (C. Moler)

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

$$\Sigma_i x_i = 1$$

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.

PageRank Tinkerings

- Convergence Criteria
  - + stop iterating when ordering converges. (Haveliwala et.al.)
- Personalization vector: $\mathbf{E} = \mathbf{e}\,\mathbf{v}^T$ $(v^T > \mathbf{0}, \sum_i v_i = 1)$
  - + does not affect convergence properties.
  - + aids against spammers.
- α
  - — for α near 1, number of iterations increases dramatically.
  - + for $\alpha \approx .85$ Google runs 50-100 iterations.
- Forced Irreducibility: $\mathbf{E}$ is one way to insure π^T exists
  - + There are others

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 buildup. (Column 4 and Pages A4 and A6)

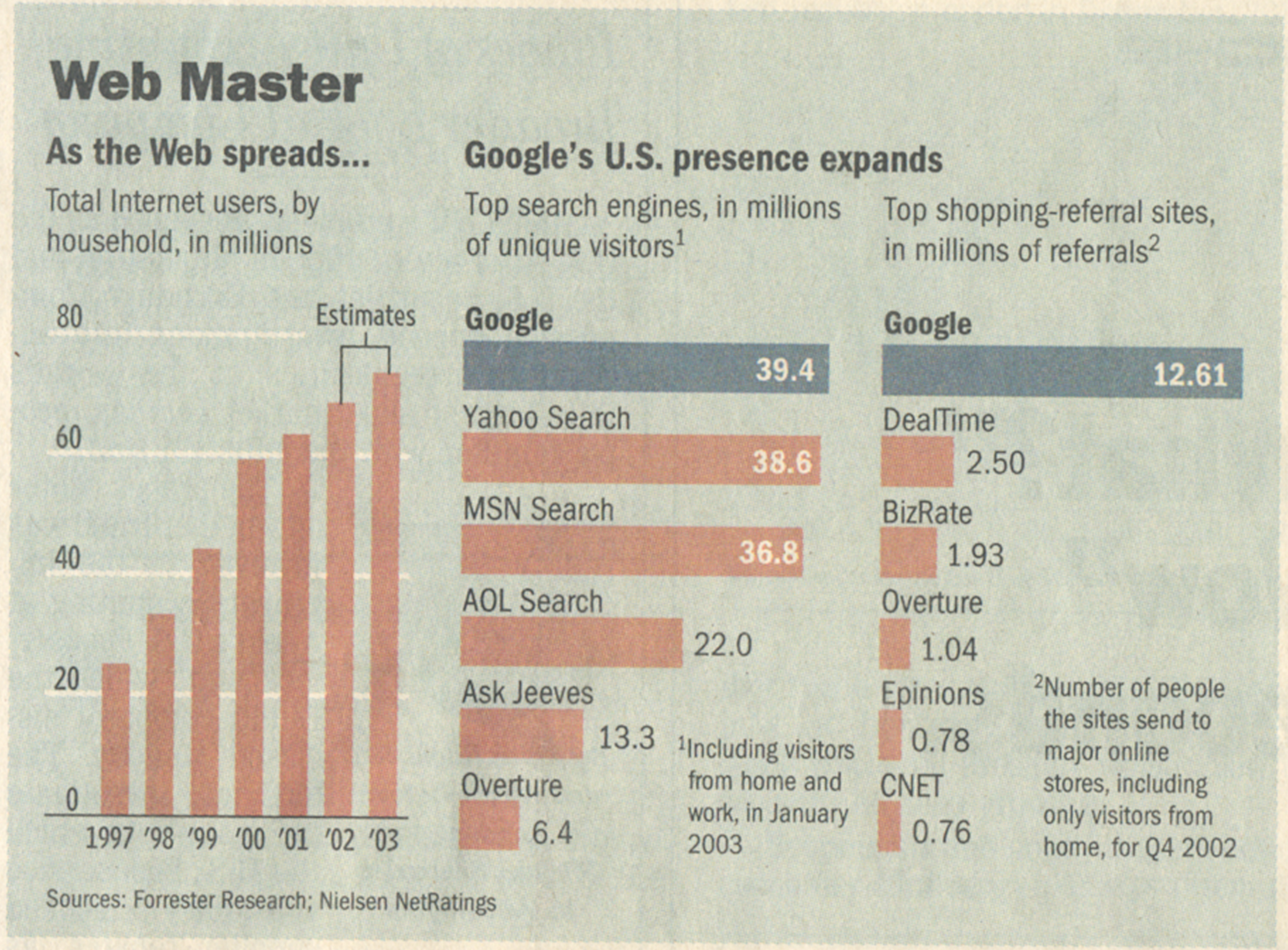
Turkey's parliament debated legislation to let the U.S. deploy 62,000 to open a northern front. Kurdish soldiers lined roads in a show of force as U.S. officials traveled into Iraq's north for an opposition conference.

* * *

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

* * *

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



[1] Including visitors from home and work, in January 2003

[2] Number of people the sites send to major online stores, including only visitors from home, for Q4 2002

Sources: Forrester Research; Nielsen NetRatings

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. Pranksters would try to

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

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

Ms. Holman, the leatherwear retailer, discovered the consequences of trying to fool Google. The 42-year-old hospital laboratory technician, who learned computer skills by troubleshooting her hospital's 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
dissatisfied customers and answe
competitive challenge from its chie
val with remodeled stores, increase
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 millio
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
a year earlier. Using comparable 13-w
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 Decen
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
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,
Fresco, later this year. But Mr. Fr
who has served as chairman since

Current Web IR Challenges

- Speed Improvements
    - identifying clusters
    - using clusters
    - enhancing power method
- Updating
- Spam Prevention