CSE326: Data Structures World Wide What?

Hannah Tang and Brian Tjaden

Summer Quarter 2002

Quick Questions

- How much wood would a woodchuck chuck if a woodchuck would chuck wood? He'd chuck as much wood as a woodchuck could if a woodchuck could chuck wood.
- Can I get web-access to our grades? Read and write permission?
```
DecreaseGrade(double amountToDecreaseYourGrade)
```
- Where can I get good ostrich meat? http://www.ostrichesonline.com/meat/meatindex.html
- When will the *dot.com* stocks recover? 21 months
- Why wasn't web invented earlier? Al Gore wasn't around
- How many nodes in the web graph? 3 billion?

More Questions

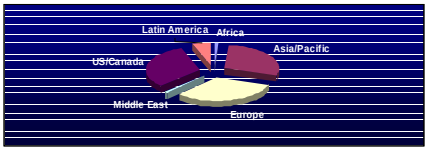
- How does a web page request (email) know where to go? How does it know how to find me?
- How do I eliminate pop-op adds?
- How much bandwidth gets used each day on the net serving web pages?
- What is the number of pages on the web that have not had their content updated in the past year?
- What happens when we run out of IP addresses? How do we keep them from colliding?
- Why don't most web routers verify the sender's IP before forwarding? Does this make them vulnerable?
- Is Microsoft's HailStorm idea of software as a service realistic?
- How do counters which count the visits (hits) to a page work? What are they used for?
- How can you connect multiple users that are accessing the same web page?
- Why can't certain ISP's access some web pages?
- How do I access accounts/web pages/etc. which are restricted?
- Why doestn't every website allow you to put "www" in front of the address?
- How does packet routing work?
- How are URL's translated into IP addresses?

Search Questions

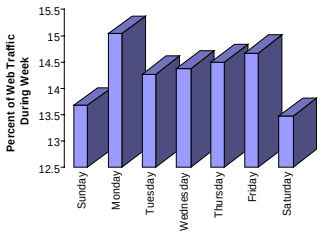
- How does web searching work?
- How do different search engines differ?
- How do they make money?
- How do they crawl and search such a big web?

Google™

More than 600 million OnLine



What's the best day to surf?



What's the most popular site?

AT WORK

	Site	Audience (in Millions)
1	Microsoft	24
2	Yahoo!	20
3	AOL Time Warner	19
4	Google	10
5	Amazon	6
6	eBay	6
7	Terra Lycos	5
8	About-Primedia	5
9	USA Network	5
10	Viacom International	3
11	CNET Networks	3
12	Excite Network	3
13	Walt Disney Internet Group	3
14	Landmark Communications	3
15	AT&T	3
16	New York Times Company	3
17	Gannett	2
18	RealNetworks	2
19	Verizon Communications	2
20	InfoSpace	2
21	The Gator Corporation	2
22	TMP Worldwide	2
23	Ask Jeeves	2
24	Adobe	2
25	Tribune Interactive	2
??	CSE326	0.00003

AT HOME

	Site	Audience (in Millions)
1	Microsoft	39
2	AOL Time Warner	37
3	Yahoo!	34
4	Google	11
5	eBay	9
6	Terra Lycos	8
7	About-Primedia	7
8	Amazon	7
9	The Gator Corporation	6
10	Viacom International	5
11	USA Network	5
12	AT&T	4
13	Excite Network	4
14	Walt Disney Internet Group	4
15	AWS Convergence Technologies	4
16	CNET Networks	4
17	eUniverse	3
18	Electronic Arts	3
19	InfoSpace	3
20	Landmark Communications	3
21	EarthLink	3
22	Classmates	3
23	Ask Jeeves	3
24	iVillage	2
25	United Online	2
??	CSE326	0.00003

Averge Internet Usage Internationally

Number of sessions per month:	18
Number of unique domains visited:	48
Page views per month:	797
Page view per surfing session:	43
Time spent per month:	9:49:53
Time spent during surfing session:	0:32:04
Duration of page view:	0:00:44

Surfing from 10,000 feet

- Type in web address, e.g., `www.amazon.com`
- DNS Lookup translates name into 32-bit IP address `207.171.181.16`
- Transmission broken up into packets (TCP/IP)
- Packets travel via internet (routers direct packets at each hop) to destination

Where does it go?

```
> traceroute www.u-tokyo.ac.jp
```

Tracing route to www.u-tokyo.ac.jp [133.11.128.254] over a maximum of 30 hops:

```
 1  <10 ms   10 ms  <10 ms  regina-GE3-1.cac.washington.edu [128.95.3.100]
 2  <10 ms  <10 ms  <10 ms  uwbr2-GE0-1.cac.washington.edu [140.142.150.24]
 3  <10 ms  <10 ms  <10 ms  prs1-wes-ge-0-0-0-0.pnw-gigapop.net [198.107.150.30]
 4  <10 ms  <10 ms  <10 ms  TRANSPAC-PWAVE.pnw-gigapop.net [198.32.170.46]
 5  110 ms  120 ms  120 ms  foundry2.otemachi.wide.ad.jp [203.178.140.216]
 6  130 ms  130 ms  140 ms  ra37-msfc-vlan16.nc.u-tokyo.ac.jp [133.11.125.121]
 7  131 ms  130 ms  140 ms  ra36-msfc-vlan3.nc.u-tokyo.ac.jp [133.11.127.66]
 8  130 ms  131 ms  140 ms  ra16-fa1-0-0.nc.u-tokyo.ac.jp [133.11.127.5]
 9  130 ms  130 ms  141 ms  foundry1.nc.u-tokyo.ac.jp [133.11.125.82]
10  130 ms  130 ms  140 ms  www.u-tokyo.ac.jp [133.11.128.254]
```

Trace complete.

Web Searching... What are we looking for?

Excite	Yahoo!
AltaVista	AlltheWeb
Lycos	Google
Metacrawler	MSN

A case study...

PageRank

- Idea! Use link structure of web to determine a web page's value
- Interpret a link form page A to page B as a vote, by page A, for page B
- Weight A's vote for B by the value of the voting page A (divided by the number of outgoing links on page A)

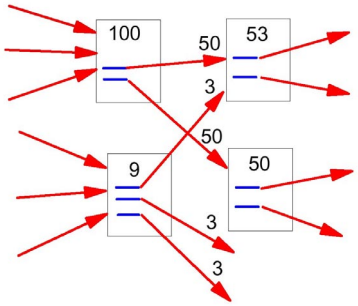
PageRank (cont.)

Assume page A has pages $T_1,T_2,...,T_n$ which point to it (i.e., are citations). Let C(B) be the number of outgoing links from a page B. Then the PageRank of page A is given by:

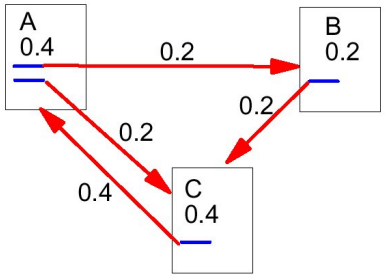
$$PR(A) = d*(PR(T_1)/C(T_1) + PR(T_2)/C(T_2) + ... + PR(T_n)/C(T_n))$$

If we view the web as a graph, where each node is a web page and and each edge is a link, then the PageRank corresponds to the principal eigenvector of the adjacency matrix and d corresponds to the principal eigenvalue.

PageRank Example



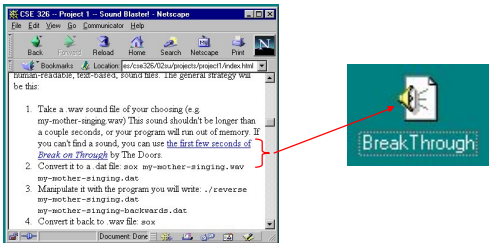
PageRank Stability



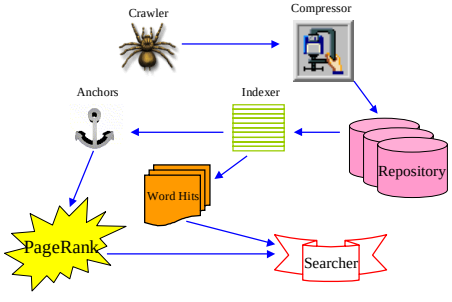
Page Info and Anchor Text

Font **Capitalization** **Count** **Position**

Associate the text of a link with the page that the link points to!



Architecture



An alternative model... Hubs and Authorities

- Use text based search to generate list of candidate pages
- Calculate hub_{score} and $authority_{score}$ for all pages in the candidate set
- Return set of pages with highest $authority_{scores}$