Search Engines:
Technology, Society, and Business

Marti Hearst

Aug 29, 2005

Today

- Discussion
- Course Goals and Logistics
- Invited Speakers and Instructors
- How the Internet / Web Works
- How Search Engines Work

A Seminar Course

- Low-key; learn something new!
- Both undergrads and graduate students.
- Very wide-ranging backgrounds

Undergraduates

Undeclared	17
CS/EECS	11
Cogsci/Psych	4
Other eng.	4
Business	3
Interdisc. Studies	3
Math/Stat	3
Humanities	2
Double major	2

Grad students

SIMS	14
CS/EECS	8
Business	4

Course Goals

- Gain an interdisciplinary understanding of search engines and related technologies.
  - How they work
  - How they affect communication
  - How they affect business
  - How they are changing our understanding of information and knowledge.
- Make the techy parts understandable for everyone.

Class Format

- Lectures by up-to-date experts
  - The class is being webcast, but not live.
  - Some speaker may decline webcasting.
- Discussion sections to explore issues more deeply
- A few short homework assignments, turned in online
- A paper or project on a topic of your choice
    - Topics will need to be approved by TAs/Prof

Class Attendance

- You *must* attend class.
  - We want a good audience for our fantastic speakers.
  - Counting today, there are 14 lectures.
  - You can miss only one class. Each class missed beyond that will be a reduction of one letter grade.
  - At the beginning of each class, the TAs will mark your name off a list; you must show your student ID.

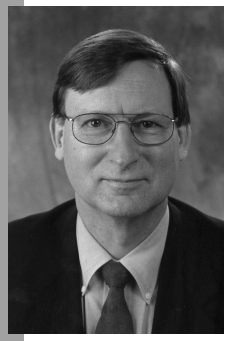
Lecturers















Instructor Background

Prof. Marti Hearst

- Associate Professor in SIMS
  - Affiliate position in the CS department
  - PhD in Computer Science from UC Berkeley
- Research areas:
  - Search, especially user interfaces for search
  - Computational linguistics
  - Information Visualization
- Industry Experience
  - Researcher at Xerox PARC for many years
  - Worked at HP, IBM
  - Now a member of the Scientific Advisory Board for Yahoo! search

TAs

- Fredrik Wallenberg
  - SIMS PhD student

- Helen Kim
  - SIMS Masters Student

- Office hours TBD

What is SIMS?

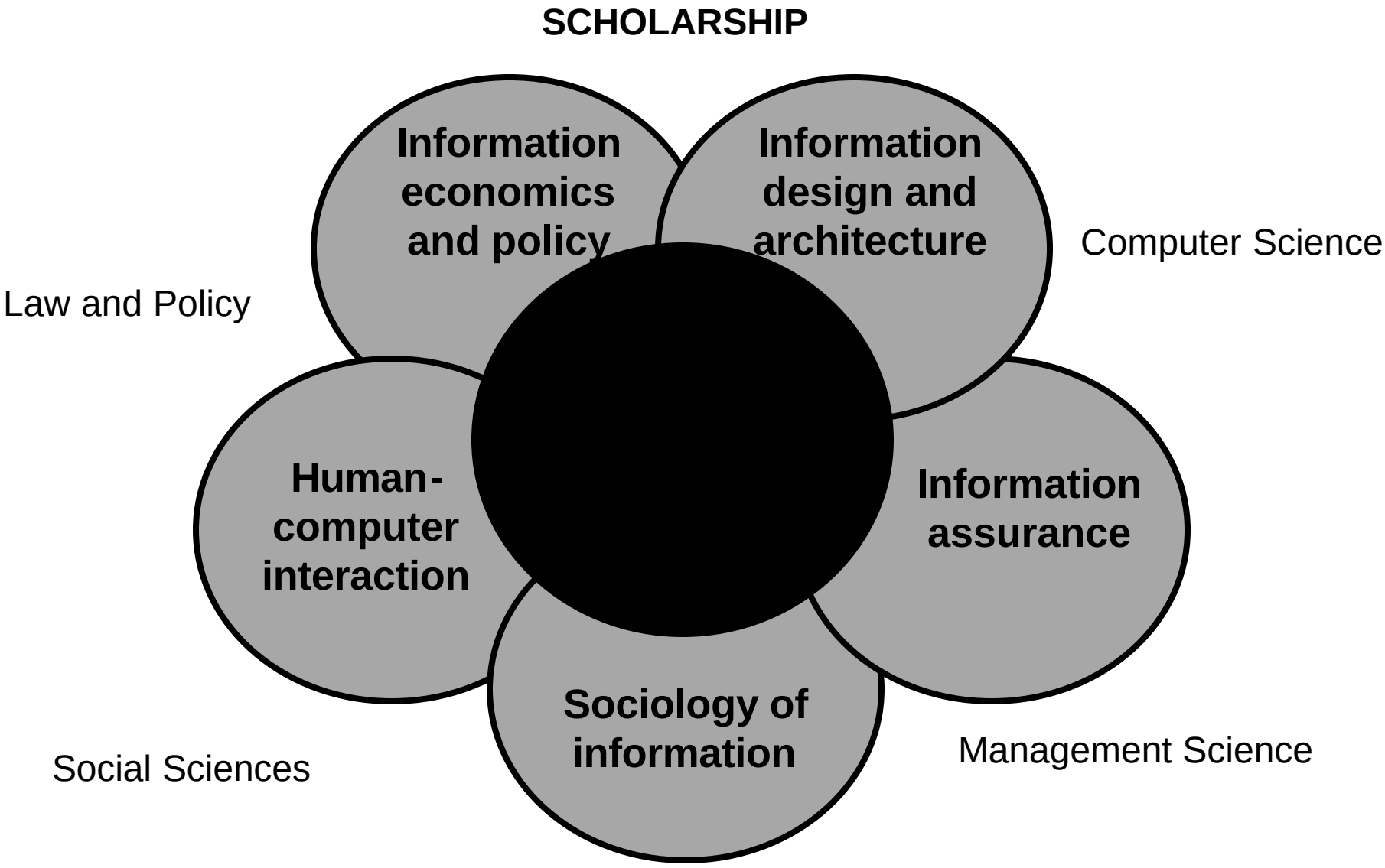
- School of Information Management & Systems
- Newest school on campus; started in 1997
- We have a PhD program and a professional masters degree
  - Like MBAs and Journalism school
- Faculty have diverse backgrounds
  - Computer science, economics, law, political science, sociology, and others.

SIMS Mission

We are developing scholars, entrepreneurs, and public leaders who can transform information into knowledge and understanding.

SIMS

SIMS


SCHOLARSHIP
Information economics and policy
Information design and architecture
Computer Science
Law and Policy
Human-computer interaction
Information assurance
Sociology of information
Social Sciences
Management Science
PROFESSIONAL SKILLS

SIMS Courses (Sample)

- Information in Society
- Database Design
- Information Visualization and Presentation
- Open Source Software: Economic, Legal & Social Implications
- Web Services
- The Quality of Information

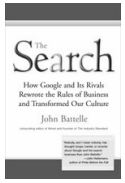
Master's student placements

Representative employers:

- **Google, eBay, Yahoo!, Microsoft, Oracle, HP**
- **UC, Kaiser, US Government, CA Digital Library**
- **Entrepreneurial**

The Next Two Weeks

- We are cancelling the Thurs 4-5pm discussion section.
  - So please switch to another if you're enrolled in it.
- No discussion section this week
  - (Aug 30-Sept 1)
  - Read Chapter 1-3 of "The Search"
- No lecture next week (campus holiday) but we will have discussion sections next week:
  - (Sept 6-8)
  - Discussion of how search engines work; learn more about HTML.
- Monday, Sept 12: John Battelle



How Search Engines Work

How Do Search Engines Work?

- Say a user named Oski using his computer at home (or in, say, Seoul) wants to find information about IS141?
- What happens when he:
  - Brings up a search engine home page?
  - Types his query?
- First we have to understand how the network works!
- Then we can understand search engines.

Internet vs. WWW

- **Internet** and **Web** are not synonymous
- Internet is a global communication network connecting millions of computers.
- World Wide Web (WWW) is one component of the Internet, along with e-mail, chat, etc.
- Now we'll talk about both.

Slide adapted from Lew & Davis

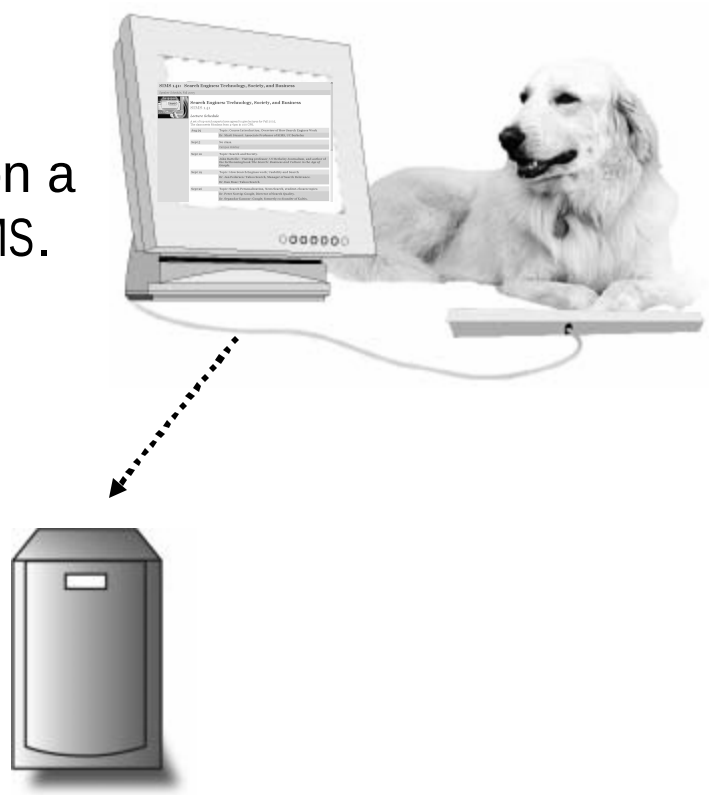
How Does the WWW Work?

- Let's say Oski received email with the address for the IS141 web page, or saw it on a flyer.
- He goes to a networked computer, and launches a web browser.
- He then types the address, known as a URL, into the address bar of the browser.
- What happens next?

(URL stands for Uniform Resource Locator)

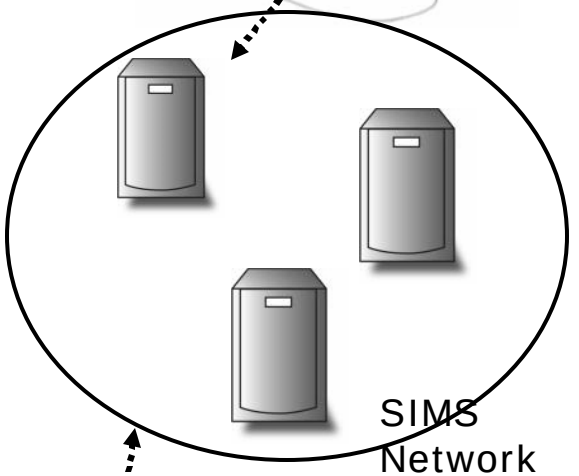
How Does the WWW Work?

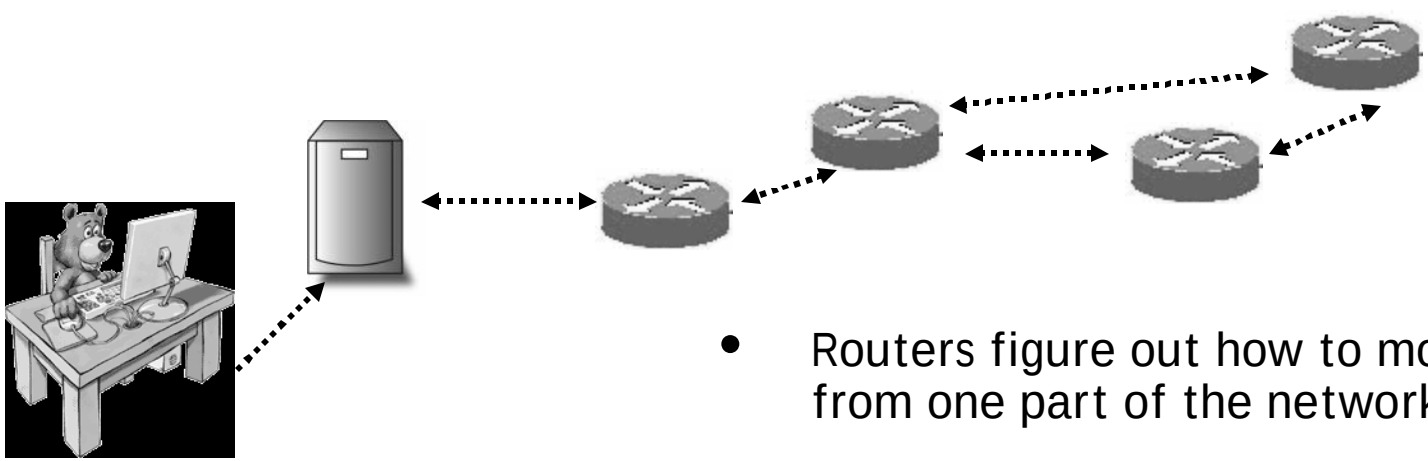
- Say Prof. Hearst has written some web pages for her class on her PC.
- She copied the pages to a directory on a computer on her local network at SIMS. The computer's name is *herald*.
- This computer is connected to the Internet and runs a program called Apache. This allows herald to act as a **web server**.



How Does the WWW Work?

- How does the computer at Oski's desk figure out where the IS 141 web pages are?
- In order for him to use the WWW, Oski's computer must be connected to another machine acting as a web server (via his ISP).
- This machine is in turn connected to other computers, some of which are **routers**.

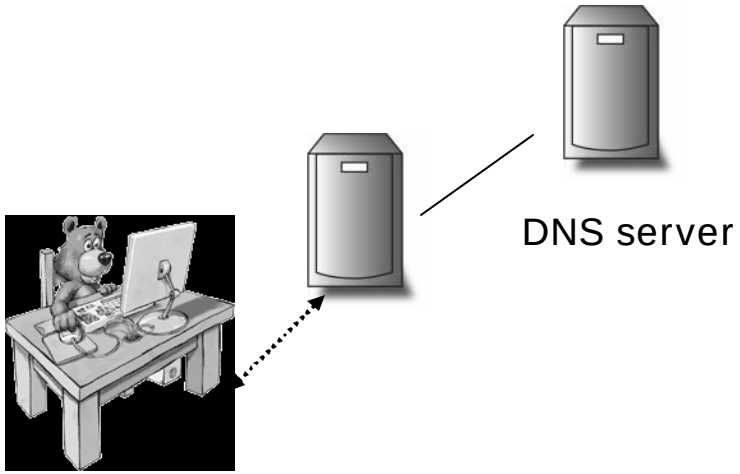




- Routers figure out how to move information from one part of the network to another.
- There are many different possible routes.

How Does the WWW Work?

- How do Oski's server and the routers know how to find the right server?
- First, the url has to be translated into a number known as an IP address.
- Oski's server connects to a Domain Names Server (DNS) that knows how to do the translation.

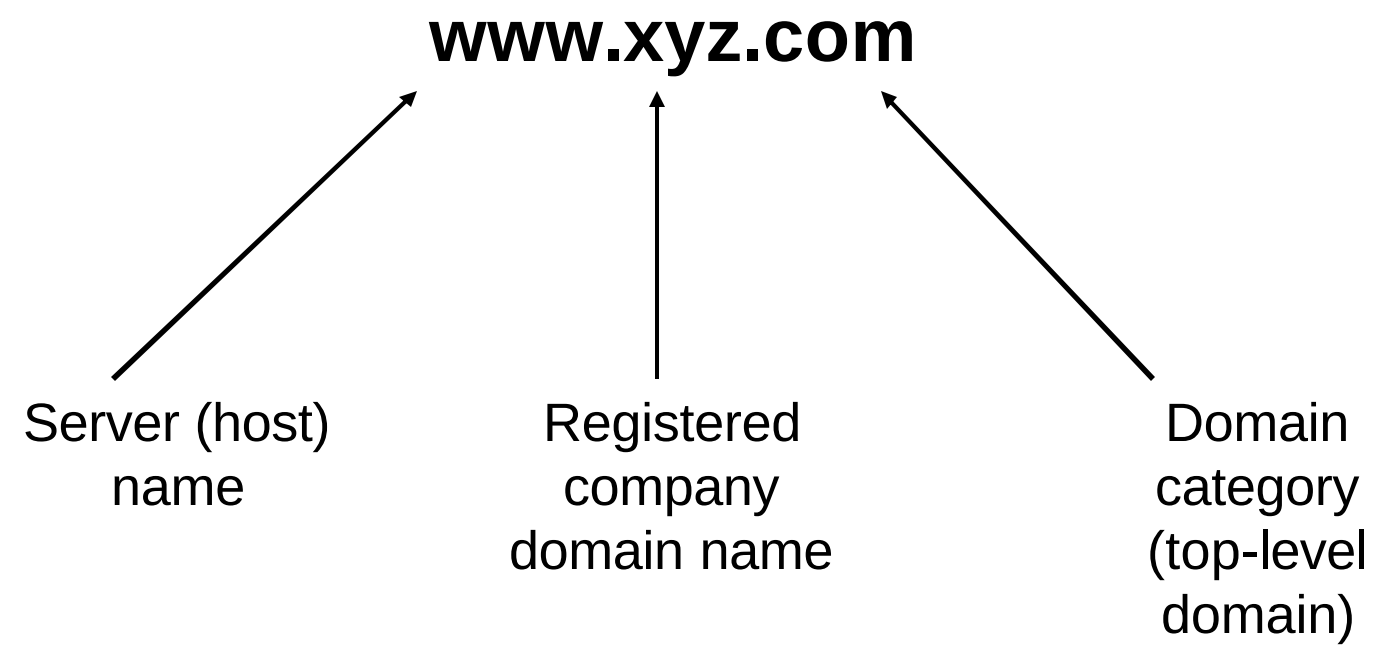


Domain Name Syntax

- Domain names are read right to left, from general to more specific locations
- For example, *www.xyz.com* can be interpreted as follows:
  - com — commercial site top-level domain
  - xyz — registered company domain name
  - www — host name (it is a convention to name web server hosts “www” which stands for “world wide web”)

Slide adapted from CIW foundations

Typical Domain Name



Domain names are part of URLs, used in web pages.

Slide adapted from CIW foundations

Top-Level Domains

- com, biz, cc — commercial or company sites
- edu — educational institutions, typically universities
- org — organizations; originally meant for clubs, associations and nonprofit groups
- mil — U.S. military
- gov — U.S. civilian government
- net — network sites, including ISPs
- int — international organizations (rarely used)

Many other top level domains are available

Slide adapted from CIW foundations

Converting Domain Names

- Domain names are for humans to read.
- The Internet actually uses numbers called **IP addresses** to describe network addresses.
- The Domain Name System (DNS) – resolves IP addresses into easily recognizable names
- For example:
  - 12.42.192.73 = *www.xyz.com*
- A domain name and its IP address refer to the same Web server.

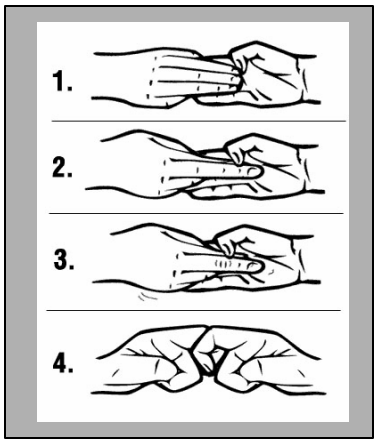
Slide adapted from CIW foundations

Internet Addresses

- The internet is a network on which each computer must have a **unique address**.
- The Internet uses **IP addresses**; for example, herald's IP address is **128.32.226.90**
- Internet Protocol version 4 (IPv4) – supports 32-bit dotted quad IP address format
  - Four sets of numbers, each set ranging from 0 to 255
  - UC Berkeley's LAN addresses range from 128.32.0.0 to 128.32.255.255
  - Other addresses in the SIMS LAN include **128.32.226.49**
- Using this setup, there are approximately 4 billion possible unique IP addresses
- Router software knows how to use the IP addresses to find the target computer.

How the Internet Works

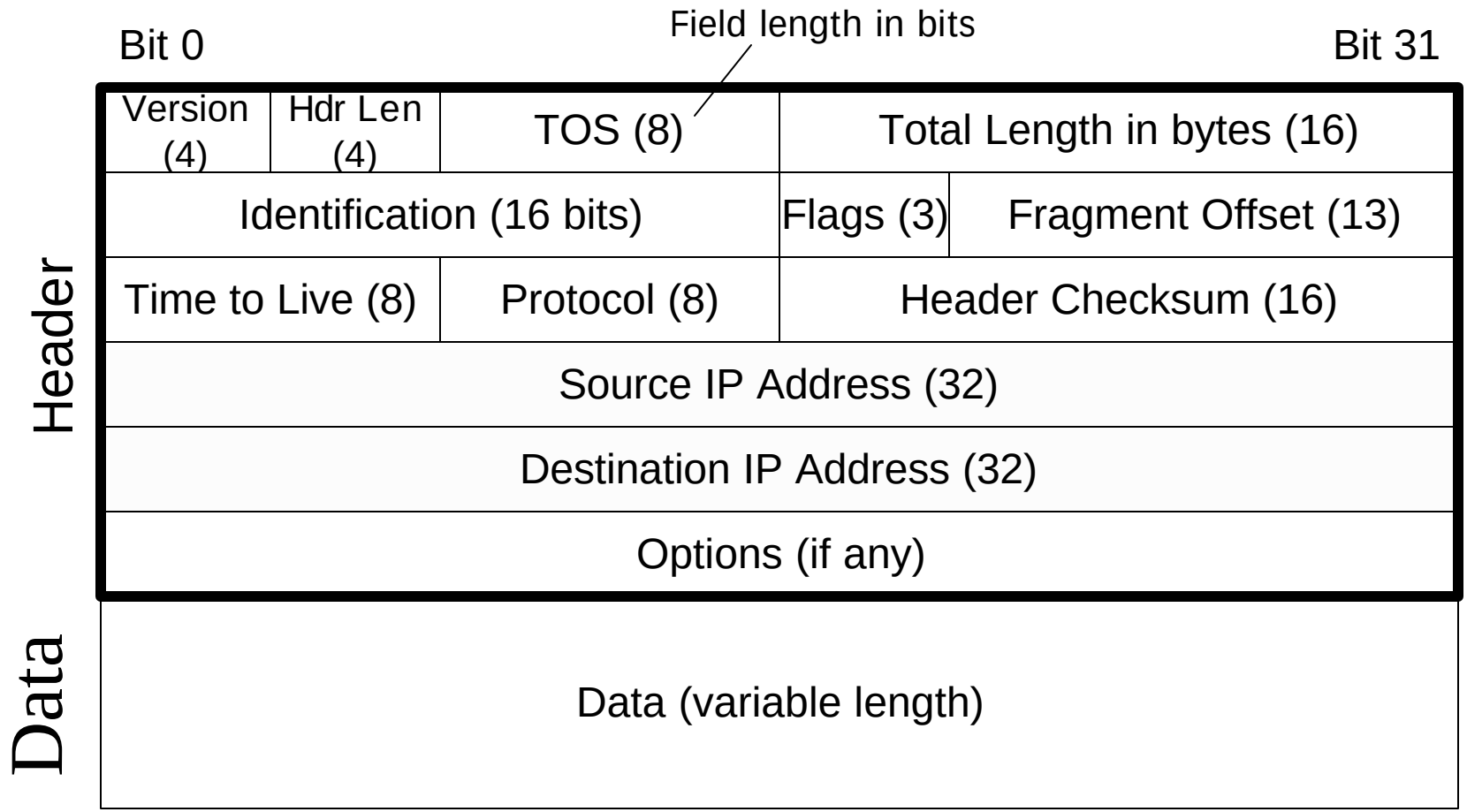
- Network Protocols:
  - Protocol – an agreed-upon format for transmitting data between two devices
    - Like a secret handshake
  - The Internet protocol is TCP/IP
  - The WWW protocol is HTTP



- Network Packets:
  - Typically a message is broken up into smaller pieces and reassembed at the receiving end.
  - These pieces of information, surrounded by address information are called **packets.**

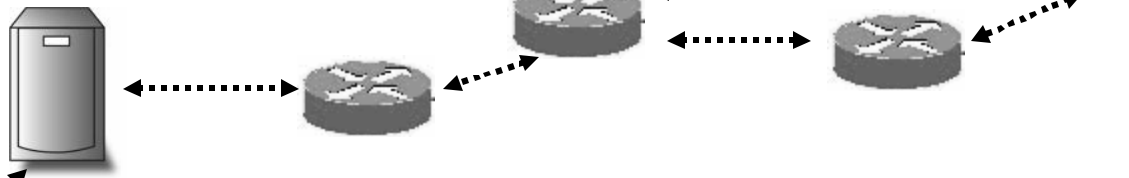
Slide adapted from CIW foundations

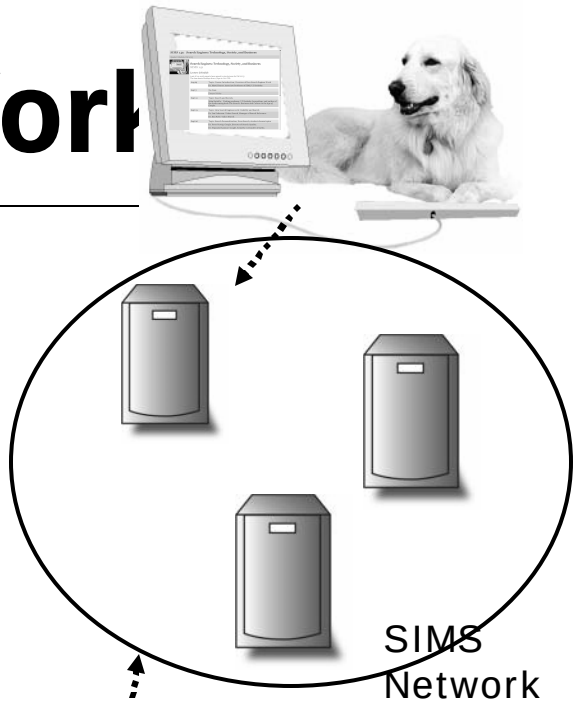
IP Packet Format (v4)



How Does the WWW Work?

- What happens now that the request for information from Oski's browser has been received by the web server *herald* at www.sims.berkeley.edu?
- The web server processes the url to figure out which page on the server is requested.
- It then sends all the information from that page back to the requesting address.





Reading a URL

http://www.sims.berkeley.edu/courses/is141/f05/index.html

http:// = HyperText Transfer Protocol

www = service name

.sims = host name

.berkeley = primary domain name

.edu/ = top level domain

courses/ = directory name

is141/f05/ = directory names

index.html = file name of web page

Slide adapted from Lew & Davis

Web Pages and HTML

- So what do we see at http://www.sims.berkeley.edu/courses/is141/f05/index.html

SIMS 141: Search Engines: Technology, Society, and Business

Course Syllabus, Fall 2005

Speaker Schedule

Search Engines: Technology, Society, and Business
SIMS 141

Mondays 4:00-6:00pm, (2 units)
2 hours of lecture per week, 1 hour of discussion per week.
CCN: 42702
Prerequisites: None.
Location: 100 Genetics & Plant Biology Bldg
Open to all undergraduate students and designed for those with little technical background.

(Graduate student version of the course)

Speaker Schedule

The organizer, Prof. Marti Hearst, is an Associate Professor at SIMS, and has done extensive research on search user interfaces. She is also on the Science Advisory Board for Search at Yahoo, Inc. She will provide the introduction to the course, devise the homework assignments, and create lectures for topics that are not covered by other speakers.

A set of top-notch experts have agreed to give lectures for Fall 2005. See the Speaker Schedule.

Sign up for Talk Announcements

Can't take the class but want to get the weekly talk announcements? Sign up here for the talk announcements list. The *only* email you will receive on this list will be the talk announcements. (To do this manually, send email to majordomo@sims.berkeley.edu with this line in the body: subscribe search-engines-talks)

Synopsis

The World Wide Web brings much of the world's knowledge into the reach of nearly everyone with a computer and an internet connection. The availability of huge quantities of information at our fingertips is transforming government, business, and many other aspects of society

Web Pages and HTML

- So what do we see at http://www.sims.berkeley.edu/courses/is141/f05/index.html
- Right-click to view the "source" or HTML code for the page.

Web Pages and HTML

- So what do we see at http://www.sims.berkeley.edu/courses/is141/f05/index.html

```
<h2>Search Engines: Technology, Society, and Business<br>
<a href=http://www.sims.berkeley.edu/courses/is141/f05/>SIMS 141</a> </h2>
<p></p>

Mondays 4:00-6:00pm, (2 units)<br>
2 hours of lecture per week, 1 hour of discussion per week.<br>
CCN: 42702<br>
Prerequisites: None. <br>
Location: <a href=http://www.berkeley.edu/map/maps/BC23.html>100 Genetics & Plant Biology Bldg</a>
Open to all undergraduate students and designed for
those with little technical background.<p></p>

(<a href=http://www.sims.berkeley.edu/courses/is290-2/f05/index.html>
Graduate student version of the course</a>)

<a href=schedule.html><h3> Speaker Schedule</h3></a>
<p></p>
The organizer, <a href=http://www.sims.berkeley.edu/~hearst>Prof. Marti Hearst</a>,
is an Associate Professor at SIMS, and has done extensive research on search user
interfaces.  She is also on the Science Advisory Board for Search at Yahoo,
Inc.  She will provide the introduction to the course, devise the homework
```

HTML

- HyperText Markup Language
  - Uses <tags> which mark up the text and tell the browser how to display the content.
  - A backslash tag means the end of the command but is sometimes optional
- Examples
  - This is **boldface text** .
  - <p> indicates a paragraph break
  - <h1> This is a large heading </h1>
  - <h3> This is a smaller heading </h3>

HTML Hyperlinks

Mondays 4:00-6:00pm, (2 units)
2 hours of lecture per week, 1 hour of discussion per w
CCN: 42702
Prerequisites: None.
Location: 100 Genetics & Plant Biology Bldg
Open to all undergraduate students and designed for t

- Hyperlink is the most important:

   100 Genetics & Plant Biology Bldg

  - The green part is called **anchor text**
    - It's the text you see on the link
  - The pink part is the url that the link will take you to if you click on it. The http:// at the front indicates the http (Web) protocol.
  - The … is the command that indicates the enclosed information is a hyperlink, and the that text between the tags is the anchor text.

- A hyperlink can be clicked on by a person OR followed by a computer program.

HTTP

- HTTP is the protocol used by the WWW
- When a user clicks on a hyperlink in their web browser, this sends an HTTP command to the Web server named in the URL
- This command usually is to “GET” the contents of the web page and return them to the user’s browser.
- It is a very simple protocol
  - It relies on the TCP/IP functionality

HTTP Request: Example

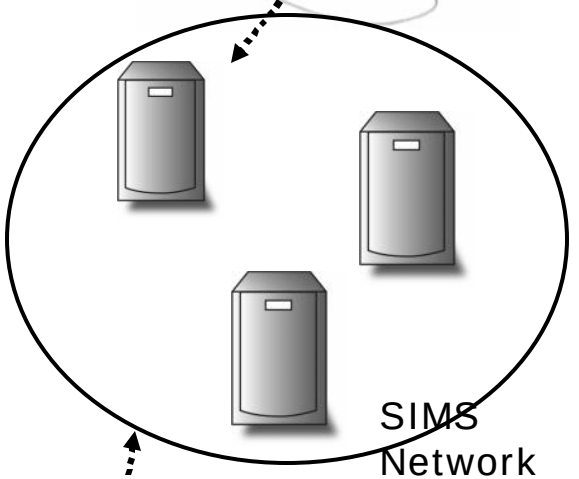
This information is received by the web server at www.sims.berkeley.edu :

Request line	GET courses/is141/s05/index.html HTTP/1.1<CRLF>
Request header	Host: www.sims.berkeley.edu <CRLF>
Blank line	<CRLF>

Because HTTP is built on TCP/IP, the web server knows which IP address to send the contents of the web page back to.

How Does the WWW Work

- When Oski typed in the url for the IS141 home page, this was turned into an HTTP request and routed to the web server in Berkeley.
- The web server then decomposed the url and figured out which web page in its directories was being asked for.
- The server then sends the HTML contents of the page back to Oski's IP address.



- Oski's browser receives these HTML contents and renders the page in graphical form.
- If he clicks on the hyperlink to the GPB map, a similar sequence of events will happen.

How the WWW/Internet Work

- More information is available online.
- There are many good glossaries:
  - http://www.alpinetech.net/glossary.html
  - http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Glossary.html
- There are good essays too:
  - http://en.wikipedia.org/wiki/Internet_Protocol
  - http://computer.howstuffworks.com/web-server.htm

How Search Engines Work

- There are MANY issues
- I’m only giving the basics today
- More will come out in future lectures

How Search Engines Work

Three main parts:

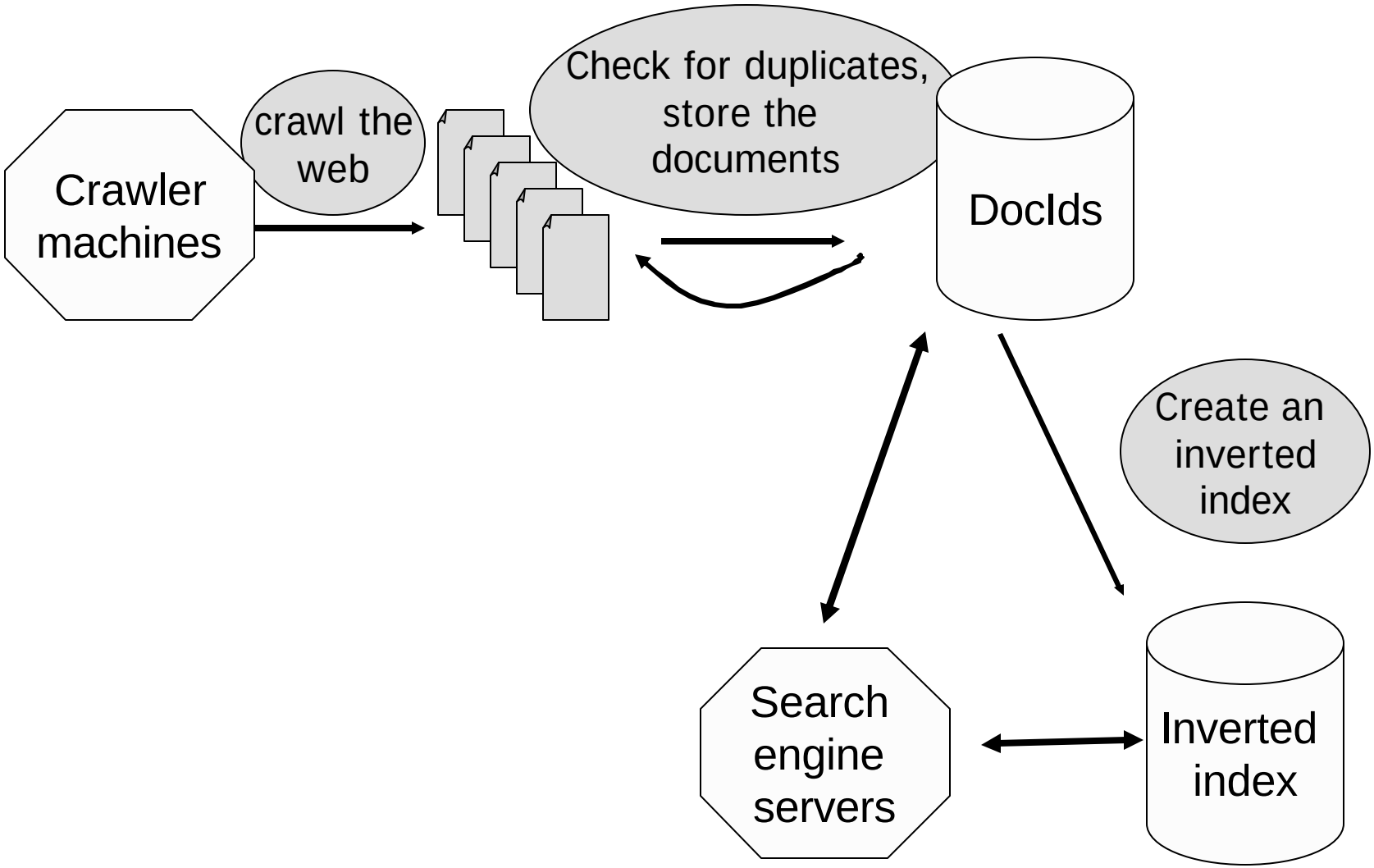
i. Gather the contents of all web pages (using a program called a **crawler** or **spider**)

ii. Organize the contents of the pages in a way that allows efficient retrieval (**indexing**)

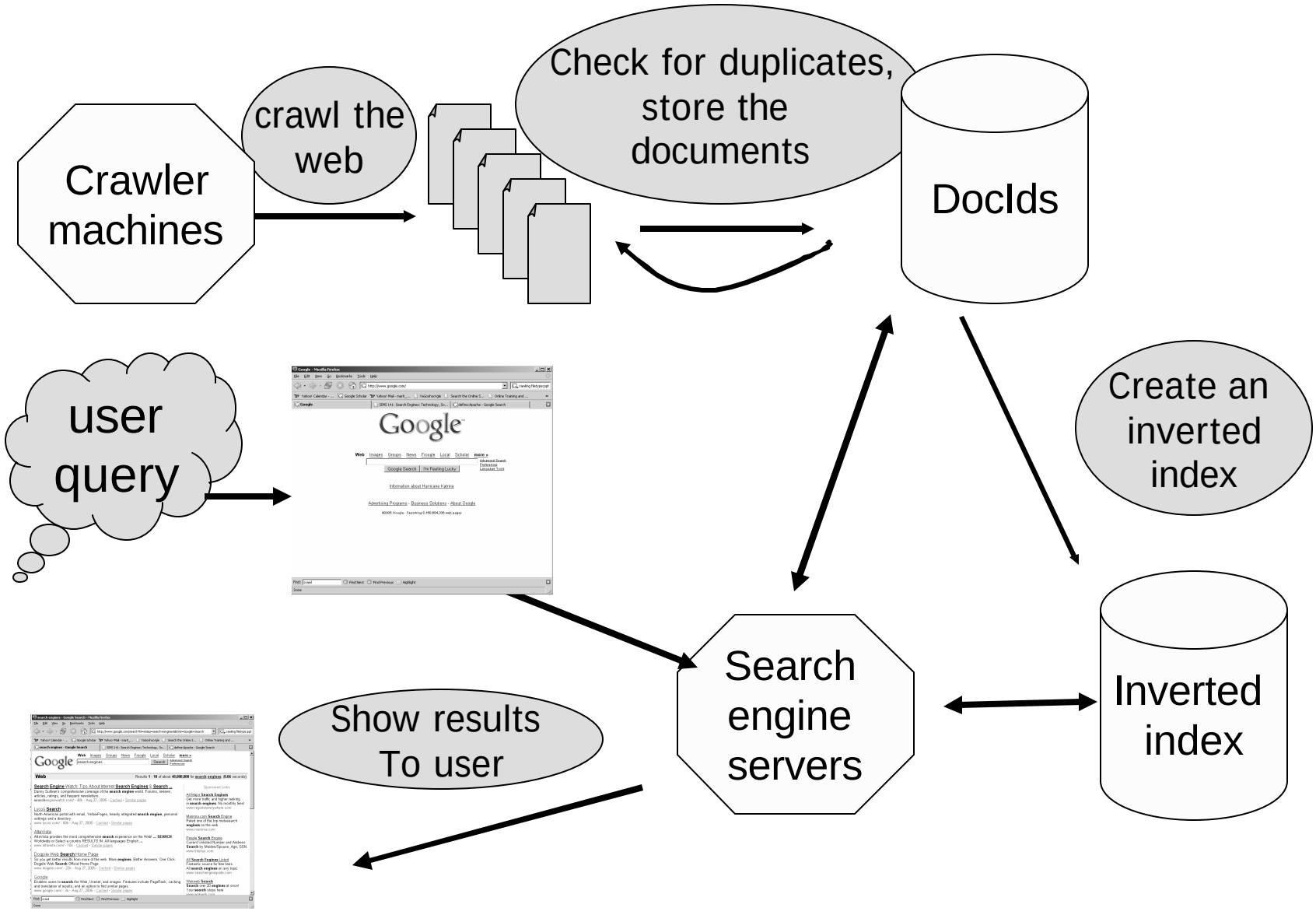
iii. Take in a query, determine which pages match, and show the results (**ranking** and **display** of results)

Slide adapted from Lew & Davis

Standard Web Search Engine Architecture

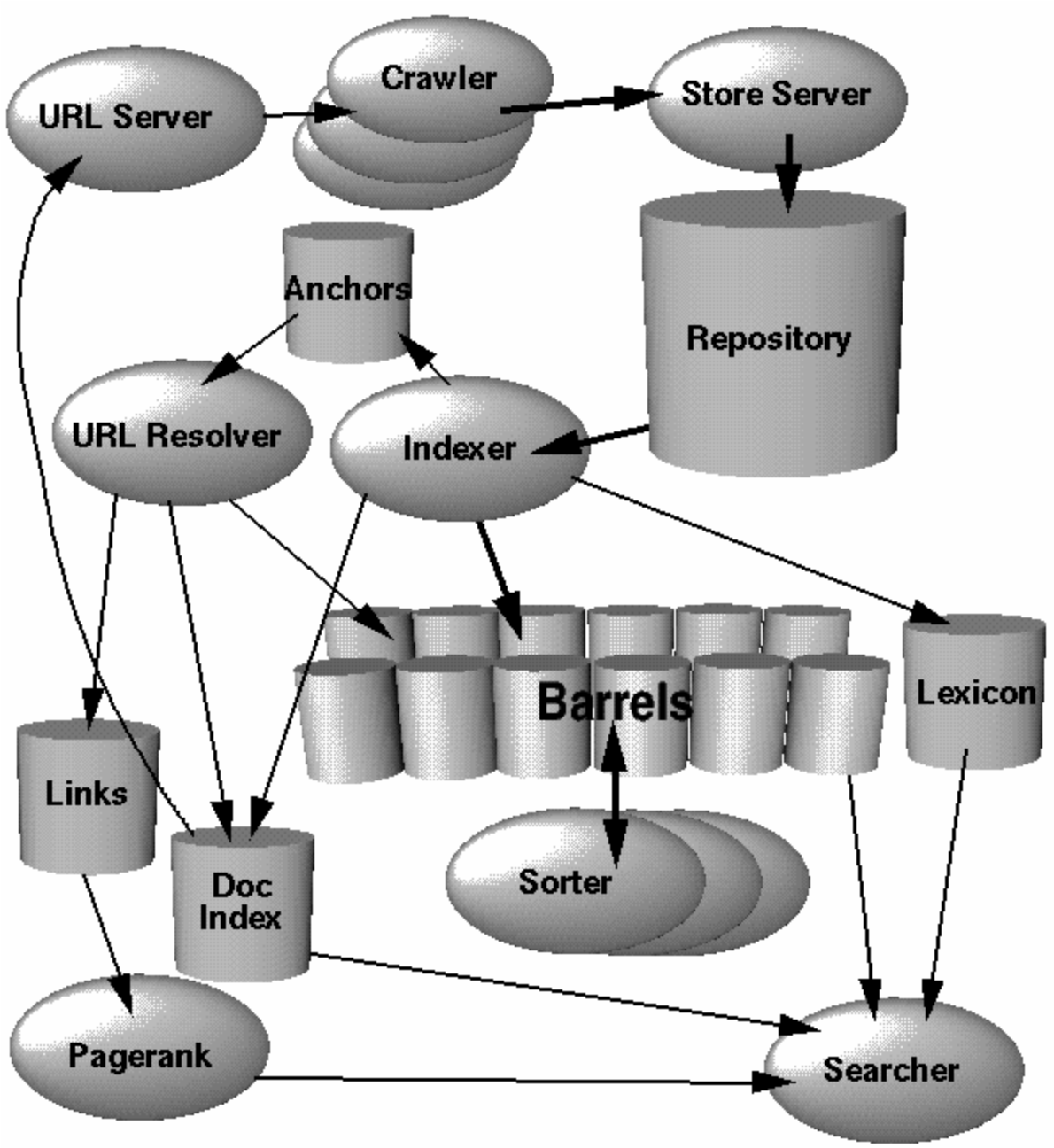


Standard Web Search Engine Architecture



More detailed architecture, from "Anatomy of a Large-Scale Hypertext Web Search Engine", Brin & Page, 1998.

http://dbpubs.stanford.edu:8090/pub/1998-8



i. Spiders or crawlers

- How to find web pages to visit and copy?
  - Can start with a list of domain names, visit the home pages there.
  - Look at the hyperlink on the home page, and follow those links to more pages.
    - Use HTTP commands to GET the pages
  - Keep a list of urls visited, and those still to be visited.
  - Each time the program loads in a new HTML page, add the links in that page to the list to be crawled.

Slide adapted from Lew & Davis

Spider behaviour varies

- Parts of a web page that are indexed
- How deeply a site is indexed
- Types of files indexed
- How frequently the site is spidered

Slide adapted from Lew & Davis

Four Laws of Crawling

- A Crawler must show identification
- A Crawler must obey the robots exclusion standard

  http://www.robotstxt.org/wc/norobots.html
- A Crawler must not hog resources
- A Crawler must report errors

Lots of tricky aspects

- Servers are often down or slow
- Hyperlinks can get the crawler into cycles
- Some websites have junk in the web pages
- Now many pages have dynamic content
  - The “hidden” web
  - E.g., schedule.berkeley.edu
    - You don’t see the course schedules until you run a query.
- The web is HUGE

The Internet Is Enormous

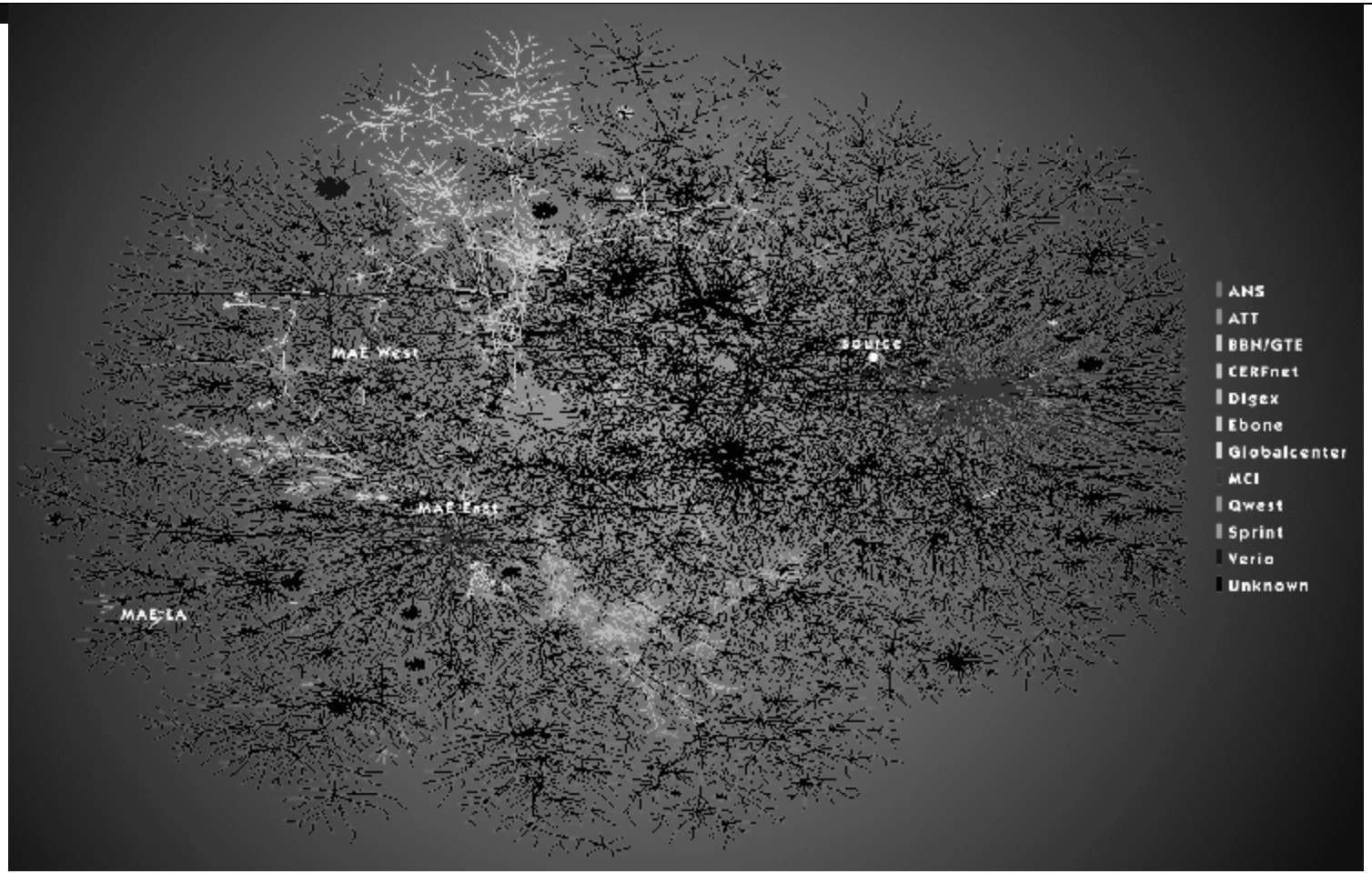


Image from http://www.nature.com/nature/webmatters/tomog/tomfigs/fig1.html

“Freshness”

- Need to keep checking pages
  - Pages change (25%,7% large changes)
    - At different frequencies
    - Who is the fastest changing?
    - Pages are removed
  - Many search engines **cache** the pages (store a copy on their own servers)

What really gets crawled?

- A small fraction of the Web that search engines know about; no search engine is exhaustive
- Not the “live” Web, but the search engine’s index
- Not the “Deep Web”
- Mostly HTML pages but other file types too: PDF, Word, PPT, etc.

Slide adapted from Lew & Davis

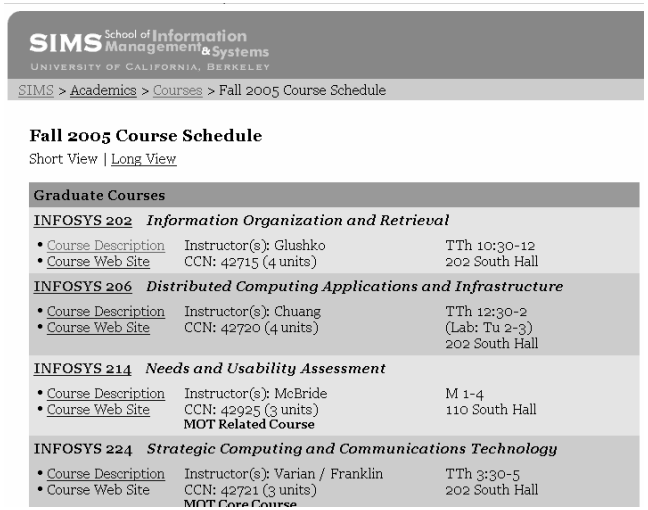
ii. Index (the database)

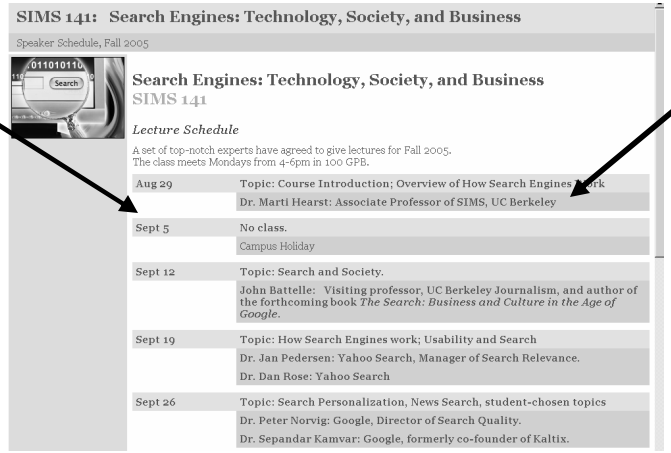
Record information about each page

- List of words
  - In the title?
  - How far down in the page?
  - Was the word in boldface?
- URLs of pages pointing to this one
- Anchor text on pages pointing to this one

Slide adapted from Lew & Davis

The importance of anchor text





INFOSYS 141

A terrific course on search engines

The anchor text summarizes what the website is about.

Inverted Index

- How to store the words for fast lookup
- Basic steps:
  - Make a “dictionary” of all the words in all of the web pages
  - For each word, list all the documents it occurs in.
  - Often omit very common words
    - **“stop words”**
  - Sometimes **stem** the words
    - (also called **morphological analysis**)
    - cats -> cat
    - running -> run

Inverted Index Example

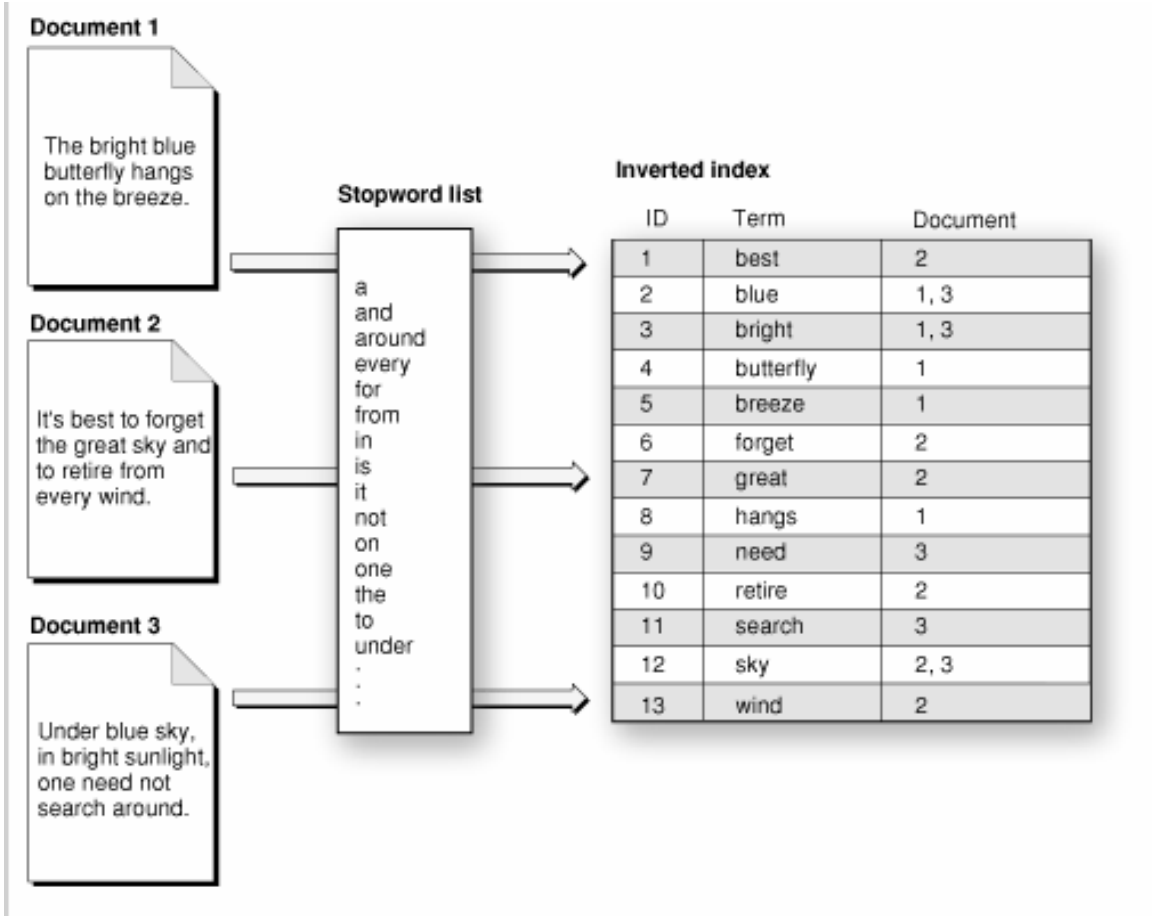
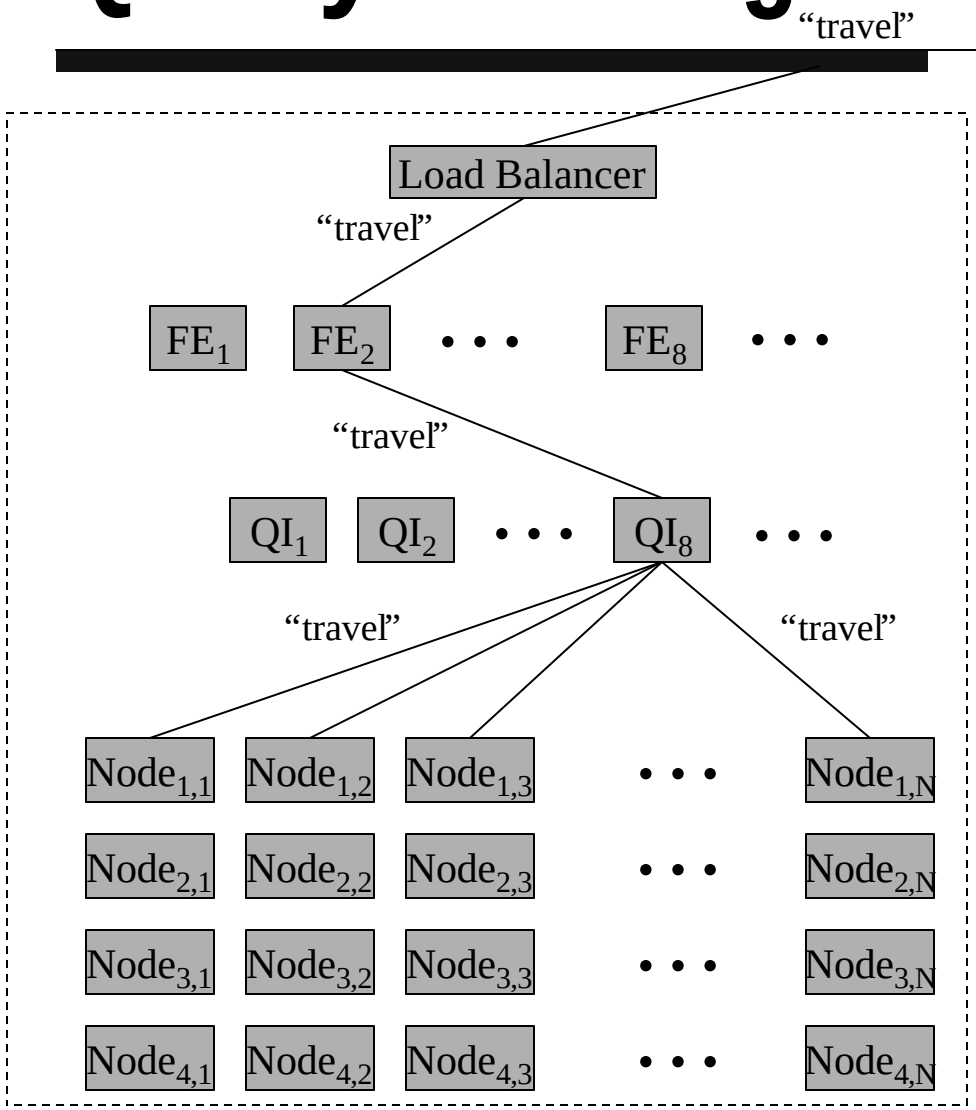


Image from http://developer.apple.com
/documentation/UserExperience/Conceptual/SearchKitConcepts/searchKit_basics/chapter_2_section_2.html

Inverted Index

- In reality, this index is HUGE
- Need to store the contents across many machines
- Need to do optimization tricks to make lookup fast.

Query Serving Architecture



- Index divided into segments each served by a node
- Each row of nodes replicated for query load
- Query integrator distributes query and merges results
- Front end creates a HTML page with the query results

iii. Results ranking

- Search engine receives a query, then
- Looks up the words in the index, retrieves many documents, then
- Rank orders the pages and extracts “snippets” or summaries containing query words.
  - Most web search engines assume the user wants all of the words (Boolean AND, not OR).
- These are complex and highly guarded algorithms unique to each search engine.

Slide adapted from Lew & Davis

Some ranking criteria

- For a given candidate result page, use:
  - Number of matching query words in the page
  - Proximity of matching words to one another
  - Location of terms within the page
  - Location of terms within tags e.g. <title>, <h1>, link text, body text
  - Anchor text on pages pointing to this one
  - Frequency of terms on the page and in general
  - Link analysis of which pages point to this one
  - (Sometimes) Click-through analysis: how often the page is clicked on
  - How “fresh” is the page
- Complex formulae combine these together.

Slide adapted from Lew & Davis

Measuring Importance of Linking

- PageRank Algorithm
  - Idea: important pages are pointed to by other important pages
  - Method:
    - Each link from one page to another is counted as a "vote" for the destination page
    - But the importance of the starting page also influences the importance of the destination page.
    - And those pages scores, in turn, depend on those linking to them.

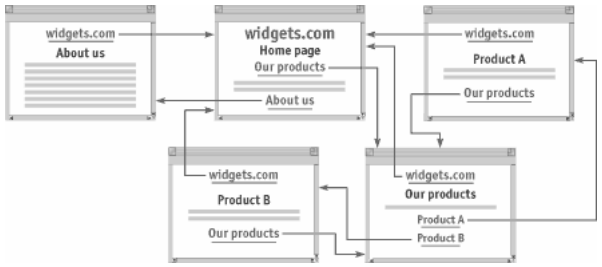


Image and explanation from http://www.economist.com/science/tq/displayStory.cfm?story_id=3172188

Measuring Importance of Linking

- Example: each page starts with 100 points.
- Each page's score is recalculated by adding up the score from each incoming link.
  - This is the score of the linking page divided by the number of outgoing links it has.
  - E.g, the page in green has 2 outgoing links and so its "points" are shared evenly by the 2 pages it links to.
- Keep repeating the score updates until no more changes.

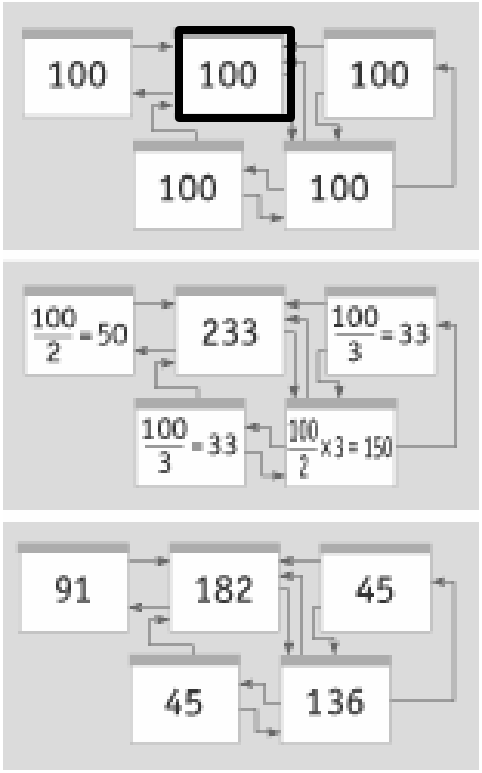


Image and explanation from http://www.economist.com/science/tq/displayStory.cfm?story_id=3172188

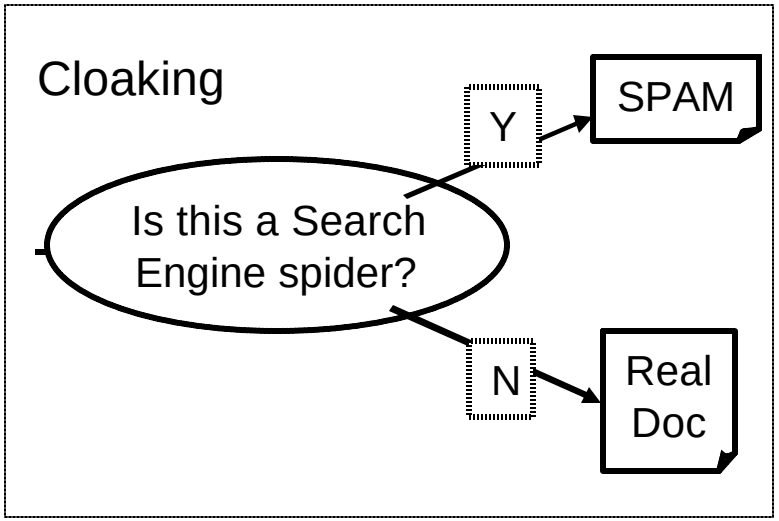
Manipulating Ranking

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forum
  - Web master world (www.webmasterworld.com)

Slide adapted from Manning, Raghavan, & Schuetze

A few spam technologies

- **Cloaking**
  - Serve fake content to search engine robot
  - *DNS cloaking:* Switch IP address. Impersonate
- **Doorway pages**
  - Pages optimized for a single keyword that re-direct to the real target page
- **Keyword Spam**
  - Misleading meta-keywords, excessive repetition of a term, fake "anchor text"
  - Hidden text with colors, CSS tricks, etc.
- **Link spamming**
  - Mutual admiration societies, hidden links, awards
  - *Domain flooding:* numerous domains that point or re-direct to a target page
- **Robots**
  - Fake click stream
  - Fake query stream
  - Millions of submissions via Add-Url



Meta-Keywords =
"… London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra, …"

Slide adapted from Manning, Raghavan, & Schuetze

Paid ranking

Pay-for-inclusion

- Deeper and more frequent indexing
- Sites are not distinguished in results display

Paid placement

- Keyword bidding for targeted ads
- Google's AdWords, Overture big players

Slide adapted from Lew & Davis

Know your search engine

- What is the default boolean operator? Are other operators supported?
- Does it index other file types like PDF?
- Is it case sensitive?
- Phrase searching?
- Proximity searching?
- Truncation?
- Advanced search features?

Slide adapted from Lew & Davis

Keyword search tips

- There are many books and websites that give searching tips; here are a few common ones:
  - Use unusual terms and proper names
  - Put most important terms first
  - Use phrases when possible
  - Make use of slang, industry jargon, local vernacular, acronyms
  - Be aware of country spellings and common misspellings
  - Frame your search like an answer or question
- For more, see http://www.googleguide.com/

Slide adapted from Lew & Davis

Search Engine Information

- www.searchenginewatch.com
- www.searchenginejournal.com
- www.searchengineshowdown.com
- http://battellemedia.com
- http://jeremy.zawodny.com/blog/

Class Attendance

- You *must* attend class.
  - We want a good audience for our fantastic speakers.
  - Counting today, there are 14 lectures.
  - You can miss only one class. Each class missed beyond that will be a reduction of one letter grade.
  - At the beginning of each class, the TAs will mark your name off a list; you must show your student ID.

The Next Two Weeks

- We are cancelling the Thurs 4-5pm discussion section.
  - So please switch to another if you're enrolled in it.
- No discussion section this week
  - (Aug 30-Sept 1)
  - Read Chapter 1-3 of "The Search"
- No lecture next week (campus holiday) but we will have discussion sections next week:
  - (Sept 6-8)
  - Discussion of how search engines work; learn more about HTML.
- Monday, Sept 12: John Battelle