Search Engines:
Technology, Society, and Business

Prof. Marti Hearst

Sept 24, 2007

How Search Engines Work

Three main parts:

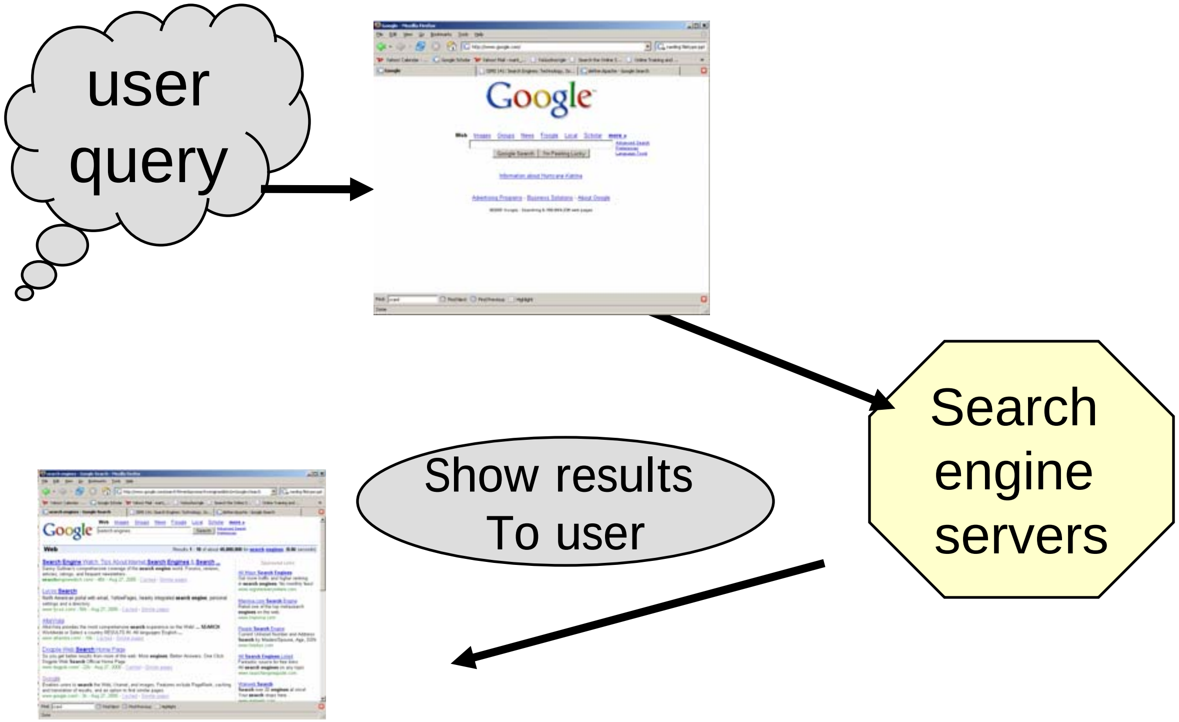
i. Gather the contents of all web pages (using a program called a **crawler** or **spider**)

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

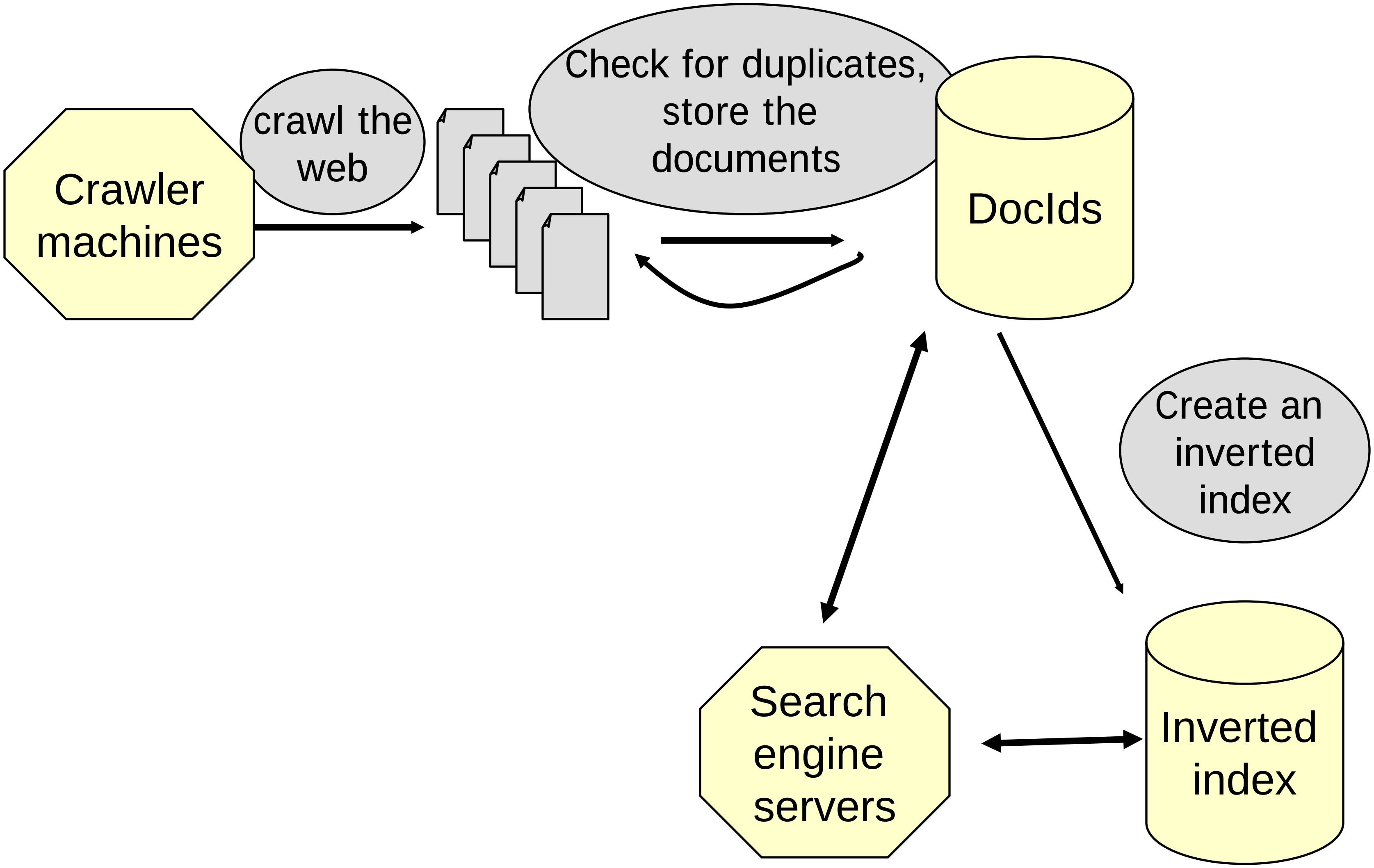


user query

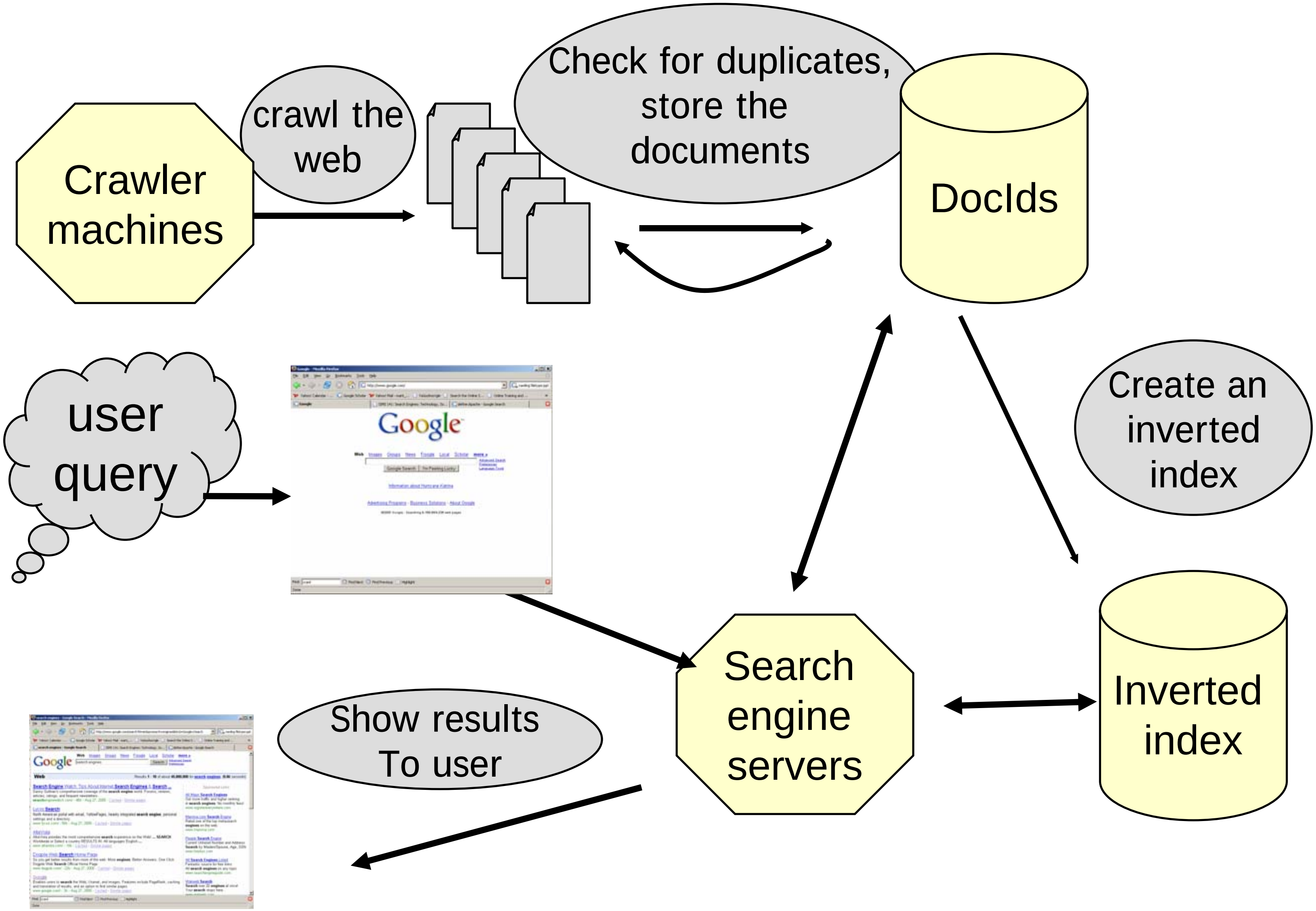
Search engine servers

Show results To user

Standard Web Search Engine Architecture



Standard Web Search Engine Architecture



i. Spiders (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

Four Laws of Crawling

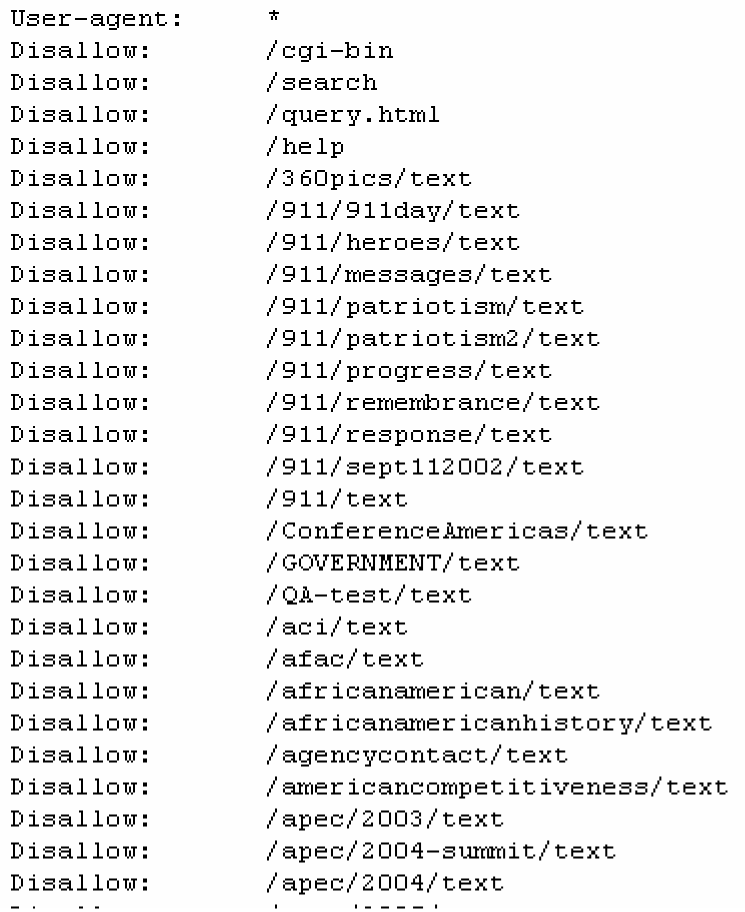
- A Crawler must show identification
- A Crawler must obey the robots exclusion standard

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

Example robots.txt file

www.whitehouse.gov/robots.txt

(just the first few lines)



```
User-agent:     *
Disallow:       /cgi-bin
Disallow:       /search
Disallow:       /query.html
Disallow:       /help
Disallow:       /360pics/text
Disallow:       /911/911day/text
Disallow:       /911/heroes/text
Disallow:       /911/messages/text
Disallow:       /911/patriotism/text
Disallow:       /911/patriotism2/text
Disallow:       /911/progress/text
Disallow:       /911/remembrance/text
Disallow:       /911/response/text
Disallow:       /911/sept112002/text
Disallow:       /911/text
Disallow:       /ConferenceAmericas/text
Disallow:       /GOVERNMENT/text
Disallow:       /QA-test/text
Disallow:       /aci/text
Disallow:       /afac/text
Disallow:       /africanamerican/text
Disallow:       /africanamericanhistory/text
Disallow:       /agencycontact/text
Disallow:       /americancompetitiveness/text
Disallow:       /apec/2003/text
Disallow:       /apec/2004-summit/text
Disallow:       /apec/2004/text
```

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

“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

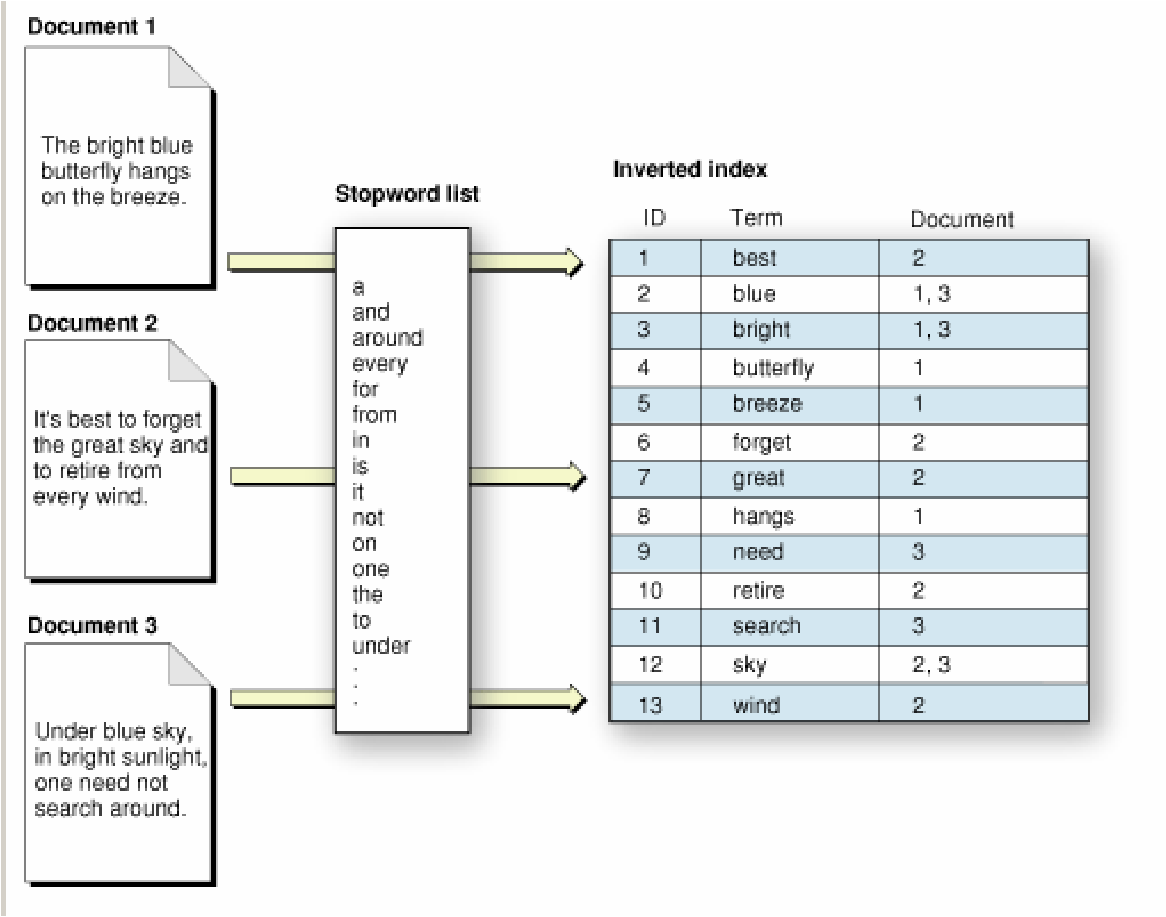
- List of words
  - In the title?
  - How far down in the page?
  - Was the word in boldface?
- URLs of pages pointing to this one
- Anchor text on pages pointing to this one

Slide adapted from Lew & Davis

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



Document 1

The bright blue butterfly hangs on the breeze.

Document 2

It's best to forget the great sky and to retire from every wind.

Document 3

Under blue sky, in bright sunlight, one need not search around.

Stopword list

a
and
around
every
for
from
in
is
it
not
on
one
the
to
under
.
.
.

Inverted index

ID	Term	Document
1	best	2
2	blue	1, 3
3	bright	1, 3
4	butterfly	1
5	breeze	1
6	forget	2
7	great	2
8	hangs	1
9	need	3
10	retire	2
11	search	3
12	sky	2, 3
13	wind	2

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.

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

Machine Learned Ranking

- Goal: Automatically construct a ranking function
  - Input:
    - Large number training examples
    - Features that predict relevance
    - Relevance metrics
  - Output:
    - Ranking function

- Enables rapid experimental cycle
  - Scientific investigation of
    - Modifications to existing features
    - New feature

What is Machine Learning?

- We don’t know how to program computers to learn the way people do
- Instead, we devise algorithms that find patterns in data.

Machine Learning Example

- Devise algorithms that find patterns in data.
- Example:
  - Start with 2 classes (italian food or chinese food)
  - Show the algorithm examples of both
  - Look at the features of each
    - Ingredients
    - Cooking style
  - Figure out which features are distinct to each class, and (optionally) how frequently they occur.
  - See a new dish: try to guess which cuisine it is in.

A Toy Example

Data Examples

Chicken parmigiana: chicken, cheese, garlic, tomatoes; bake
Spaghetti w/pesto: pasta, basil, garlic, pine nuts; saute
Pizza: flour, tomatoes, garlic, ham; bake

Kung Pao chicken: chicken, chili peppers, garlic, rice; saute
Rice noodles with shrimp: shrimp, peppers, soy, rice; saute
Pork buns: pork, onions, soy, flour; steam

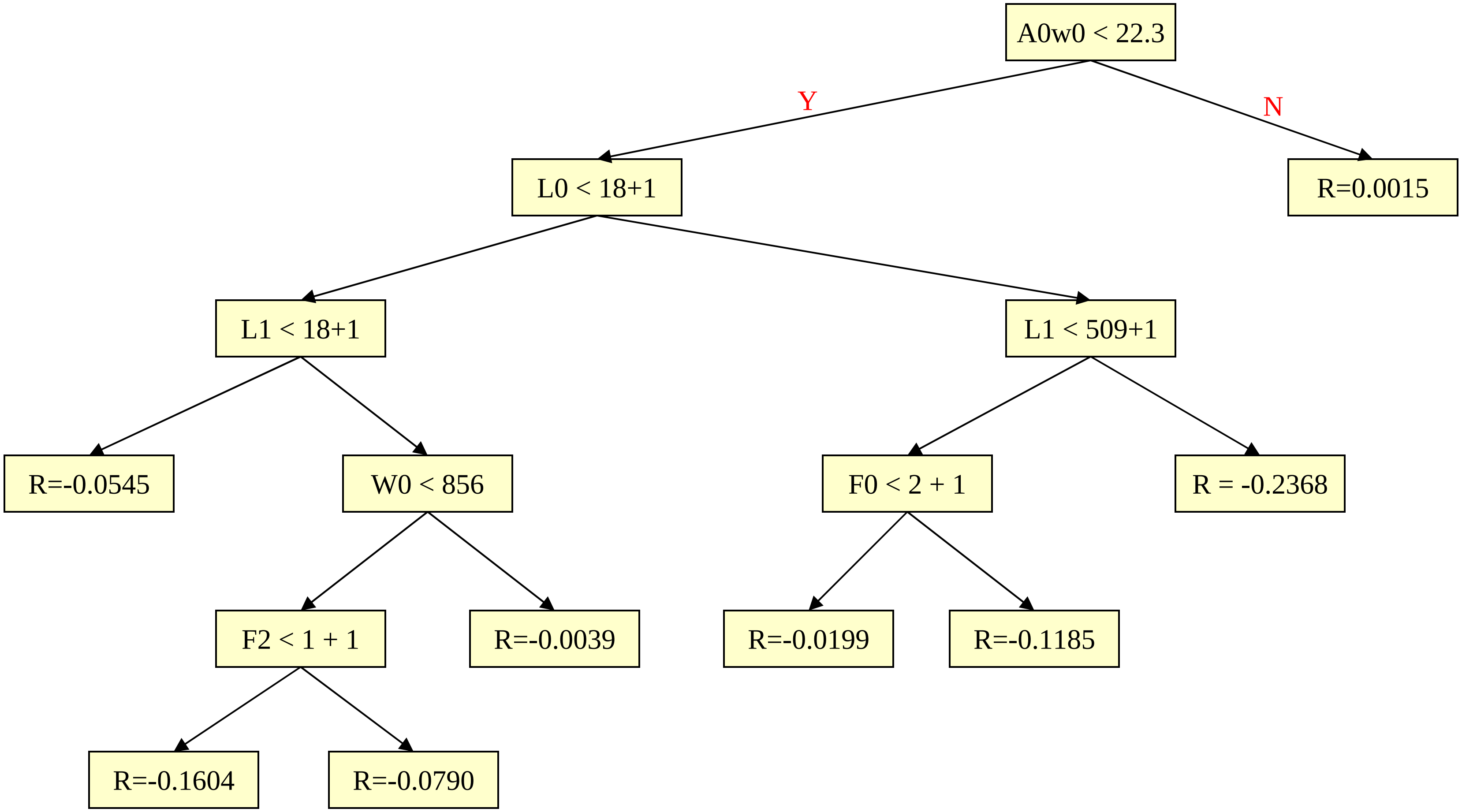
Derived Rules

```
If ingredient == tomatoes OR ingredient != rice:
          then recipe == Italian
If cooking_method == steam:
          then recipe == Chinese
```

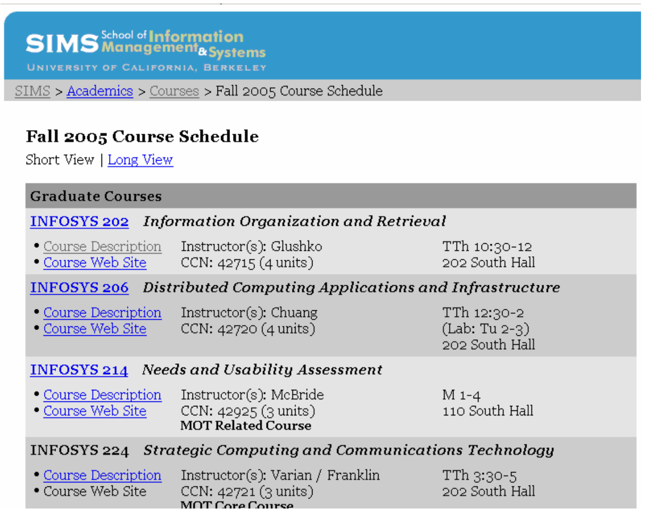
Ranking Features (from Jan Pedersen's lecture)

A0 - A4	anchor text score per term
W0 - W4	term weights
L0 - L4	first occurrence location (encodes hostname and title match)
SP	spam index: logistic regression of 85 spam filter variables (against relevance scores)
F0 - F4	term occurrence frequency within document
DCLN	document length (tokens)
ER	Eigenrank
HB	Extra-host unique inlink count
ERHB	ER*HB
A0W0 etc.	A0*W0
QA	Site factor – logistic regression of 5 site link and url count ratios
SPN	Proximity
FF	family friendly rating
UD	url depth

Ranking Decision Tree (from Jan Pedersen’s Lecture)



The importance of anchor text



SIMS School of Information Management & Systems
University of California, Berkeley

SIMS > Academics > Courses > Fall 2005 Course Schedule

Fall 2005 Course Schedule

Short View | Long View

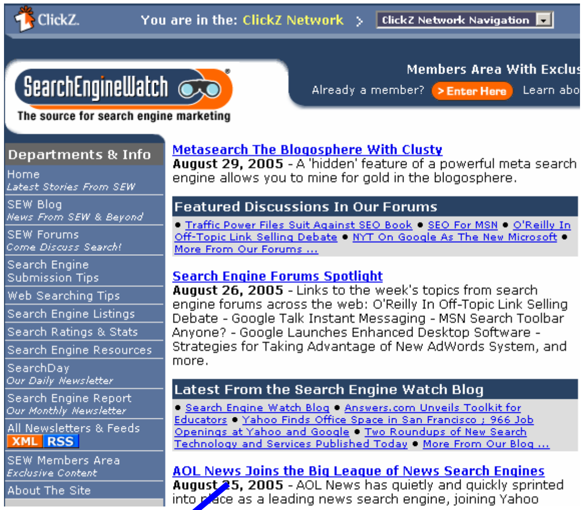
Graduate Courses

INFOSYS 202 *Information Organization and Retrieval*
- Course Description — Instructor(s): Glushko — TTh 10:30-12
- Course Web Site — CCN: 42715 (4 units) — 202 South Hall

INFOSYS 206 *Distributed Computing Applications and Infrastructure*
- Course Description — Instructor(s): Chuang — TTh 12:30-2
- Course Web Site — CCN: 42720 (4 units) — (Lab: Tu 2-3)
  202 South Hall

INFOSYS 214 *Needs and Usability Assessment*
- Course Description — Instructor(s): McBride — M 1-4
- Course Web Site — CCN: 42925 (3 units) — 110 South Hall
  MOT Related Course

INFOSYS 224 *Strategic Computing and Communications Technology*
- Course Description — Instructor(s): Varian / Franklin — TTh 3:30-5
- Course Web Site — CCN: 42721 (3 units) — 202 South Hall
  MOT Core Course



ClickZ. You are in the: ClickZ Network > ClickZ Network Navigation

SearchEngineWatch
The source for search engine marketing

Members Area With Exclus
Already a member? > Enter Here Learn abo

Departments & Info
- Home
  Latest Stories From SEW
- SEW Blog
  News From SEW & Beyond
- SEW Forums
  Come Discuss Search!
- Search Engine Submission Tips
- Web Searching Tips
- Search Engine Listings
- Search Ratings & Stats
- Search Engine Resources
- SearchDay
  Our Daily Newsletter
- Search Engine Report
  Our Monthly Newsletter
- All Newsletters & Feeds
  XML RSS
- SEW Members Area
  Exclusive Content
- About The Site

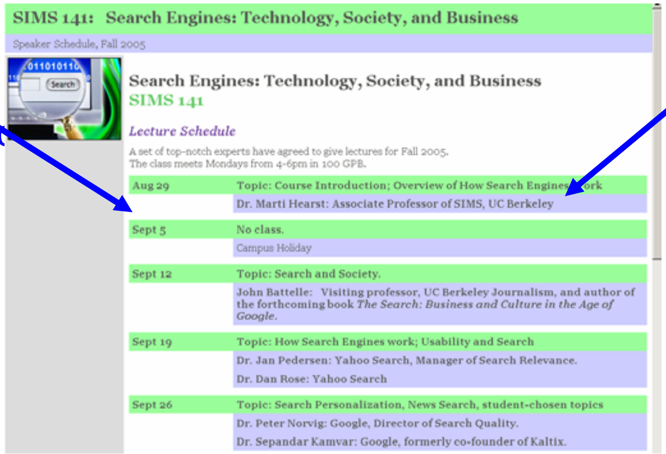
Metasearch The Blogosphere With Clusty
August 29, 2005 - A 'hidden' feature of a powerful meta search engine allows you to mine for gold in the blogosphere.

Featured Discussions In Our Forums
• Traffic Power Files Suit Against SEO Book • SEO For MSN • O'Reilly In Off-Topic Link Selling Debate • NYT On Google As The New Microsoft • More From Our Forums ...

Search Engine Forums Spotlight
August 26, 2005 - Links to the week's topics from search engine forums across the web: O'Reilly In Off-Topic Link Selling Debate - Google Talk Instant Messaging - MSN Search Toolbar Anyone? - Google Launches Enhanced Desktop Software - Strategies for Taking Advantage of New AdWords System, and more.

Latest From the Search Engine Watch Blog
• Search Engine Watch Blog • Answers.com Unveils Toolkit for Educators • Yahoo Finds Office Space in San Francisco ; 966 Job Openings at Yahoo and Google • Two Roundups of New Search Technology and Services Published Today • More From Our Blog ...

AOL News Joins the Big League of News Search Engines
August 25, 2005 - AOL News has quietly and quickly sprinted into place as a leading news search engine, joining Yahoo



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

Speaker Schedule, Fall 2005

Search Engines: Technology, Society, and Business
SIMS 141

Lecture Schedule

A set of top-notch experts have agreed to give lectures for Fall 2005.
The class meets Mondays from 4-6pm in 100 GPB.

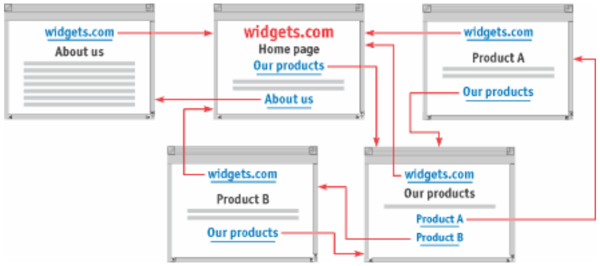
Aug 29	Topic: Course Introduction; Overview of How Search Engines Work
	Dr. Marti Hearst: Associate Professor of SIMS, UC Berkeley
Sept 5	No class.
	Campus Holiday
Sept 12	Topic: Search and Society.
	John Battelle: Visiting professor, UC Berkeley Journalism, and author of the forthcoming book *The Search: Business and Culture in the Age of Google.*
Sept 19	Topic: How Search Engines work; Usability and Search
	Dr. Jan Pedersen: Yahoo Search, Manager of Search Relevance.
	Dr. Dan Rose: Yahoo Search
Sept 26	Topic: Search Personalization, News Search, student-chosen topics
	Dr. Peter Norvig: Google, Director of Search Quality.
	Dr. Sepandar Kamvar: Google, formerly co-founder of Kaltix.

<a href=http://courses.ischoo
i141 </a>

<a href=http://courses.ischool...>
A terrific course on search engines </a>

The anchor text summarizes what the website is about.

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.

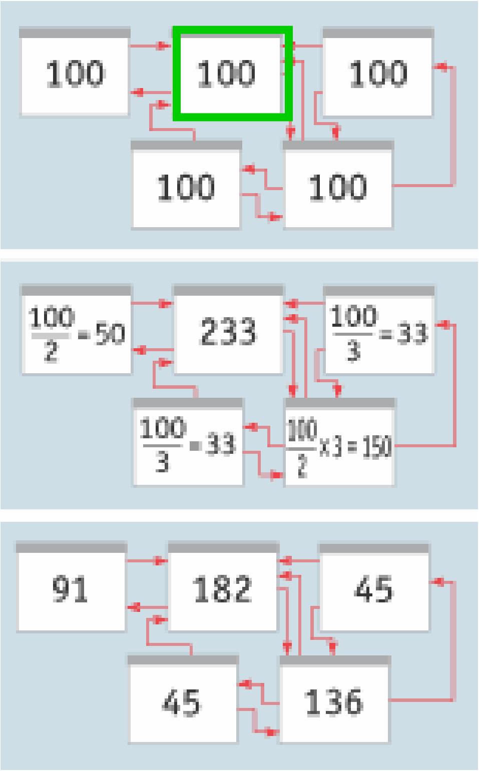


Image and explanation from http://www.economist.com/science/tq/displayStory.cfm?story_id=3172188

Class Exercise

- Students as web pages and a search engine
- Web pages:
  - Web site = where you live
  - Hyperlinks = who you know in class
  - Web page = Beatle’s song title

Unit 2

Jane
Tran

I
Wanna
Hold
Your
hand

Class Exercise

- Crawlers: follow the links between web pages
- Indexers: record information about each document
- Ranking algorithm: compute which documents to retrieve, and their order
- Human: search the web!

Crawler

- Get the first page (student) (from a pre-defined list).
- Write down the other students that this student links to (the people hyperlinks)
- Assign each document (student) a unique ID (number)
- Visit each of these in turn
- Be sure to eliminate duplicates!

Indexers

- Record the following information
- Write down each word that appears in the document
- Write down also the ID of that document (student)
- If you’ve seen that word before, add this document to that word’s list of document IDs

Ranking Algorithm

- For a given query:
  - Ask the indexers to tell it the document IDs that contain those words
  - Compute a score based on:
    - How often the words of the query occur in the document (if the word falls in the doc multiple times, that is better)
    - How popular the web site (student housing location) is.
    - How long the document is (shorter is better)
  - Formula:
    - Score for a document =
      - # hits in query + #pages in site – length(document)
  - List the results in sorted order.

Test Your Understanding

- What is the difference between the WWW and the Internet?

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

Test Your Understanding

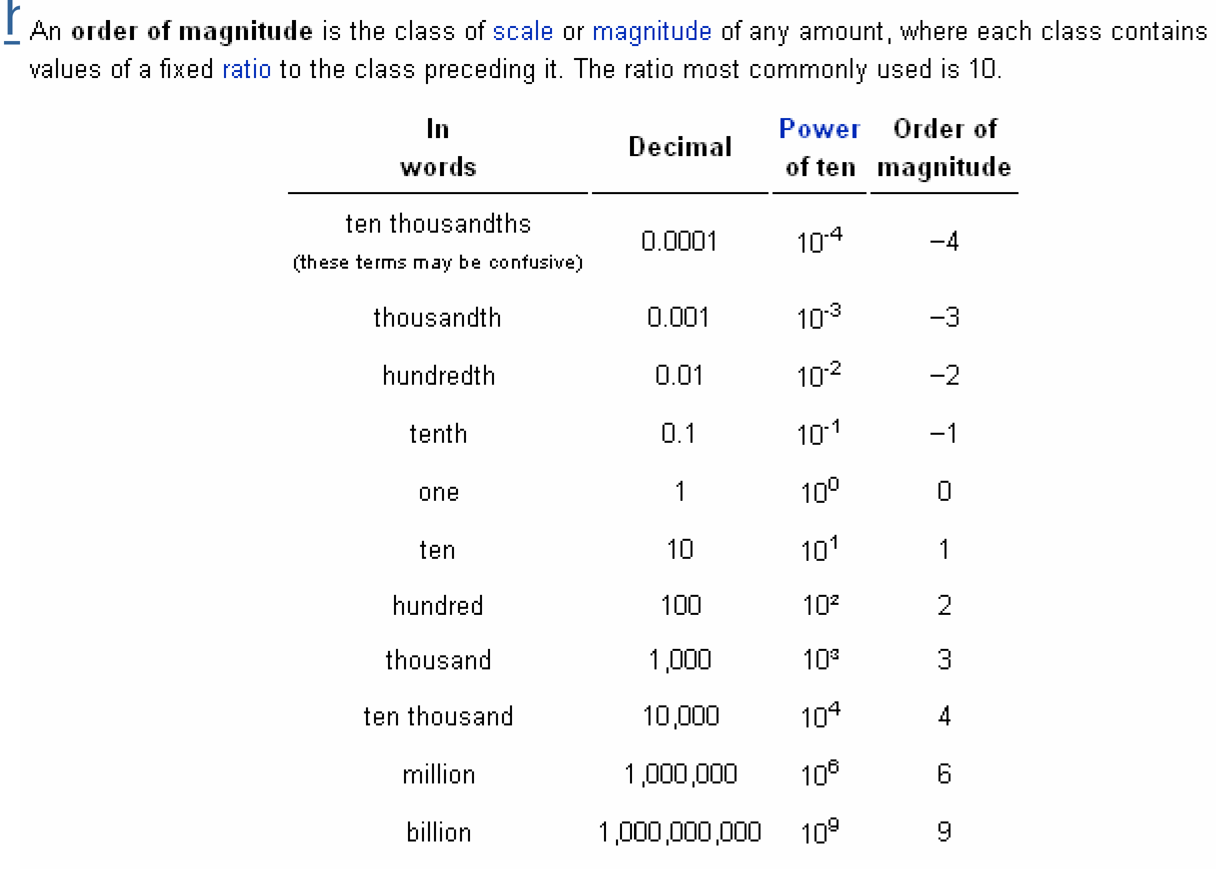
- How many queries are there per day to a major search engine?
- How much data is in the index of a major search engine?
- How many computers act as servers for a major search engine?

Test Your Understanding

- How many queries are there per day to a major search engine?
  - Hundreds of millions (NYTimes article)
- How much data is in the index of a major search engine?
  - Billions of documents
  - Petabytes of data
- How many computers act as servers for a major search engine?
  - Hundreds of thousands, maybe millions

What is a Petabyte?
Start with Orders of Magnitude

http://micro.magnet.fsu.edu/primer/java/scienceopticsu/powersof10/index.html

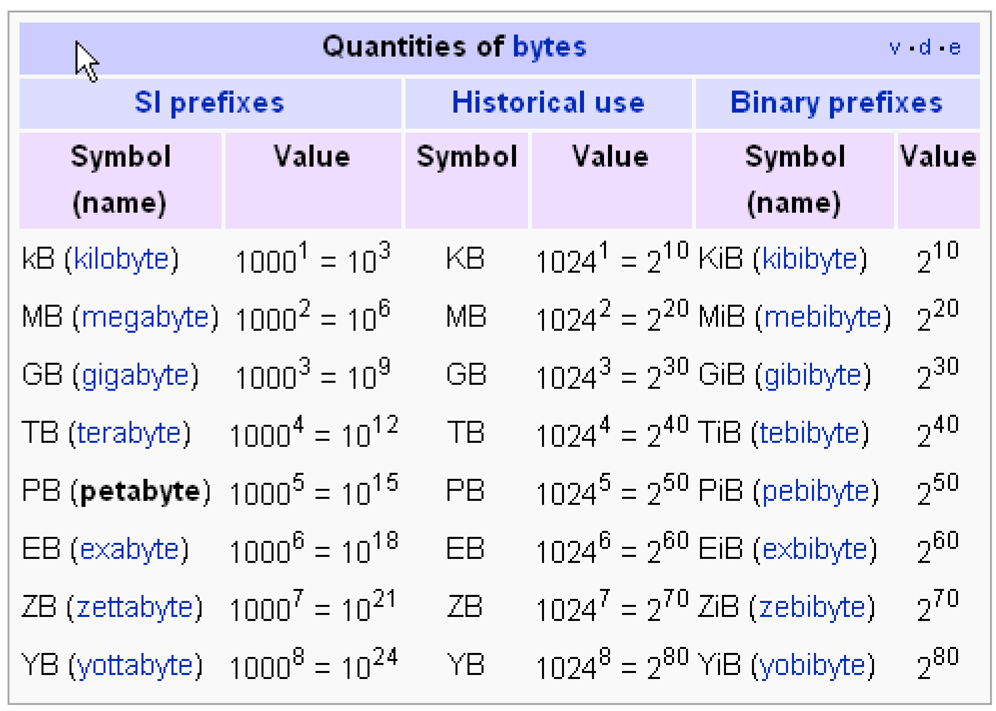


An **order of magnitude** is the class of scale or magnitude of any amount, where each class contains values of a fixed ratio to the class preceding it. The ratio most commonly used is 10.

In words	Decimal	Power of ten	Order of magnitude
ten thousandths (these terms may be confusive)	0.0001	10^{-4}	−4
thousandth	0.001	10^{-3}	−3
hundredth	0.01	10^{-2}	−2
tenth	0.1	10^{-1}	−1
one	1	10^{0}	0
ten	10	10^{1}	1
hundred	100	10^{2}	2
thousand	1,000	10^{3}	3
ten thousand	10,000	10^{4}	4
million	1,000,000	10^{6}	6
billion	1,000,000,000	10^{9}	9

What is a Petabyte?

It is 10 million gigabytes



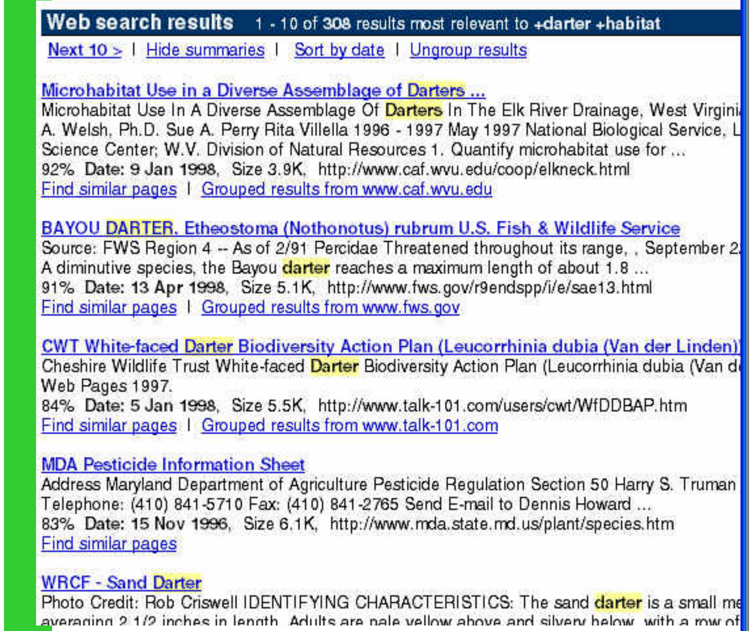
Quantities of bytes					v · d · e
SI prefixes		Historical use		Binary prefixes	
Symbol (name)	Value	Symbol	Value	Symbol (name)	Value
kB (kilobyte)	$1000^1 = 10^3$	KB	$1024^1 = 2^{10}$	KiB (kibibyte)	2^{10}
MB (megabyte)	$1000^2 = 10^6$	MB	$1024^2 = 2^{20}$	MiB (mebibyte)	2^{20}
GB (gigabyte)	$1000^3 = 10^9$	GB	$1024^3 = 2^{30}$	GiB (gibibyte)	2^{30}
TB (terabyte)	$1000^4 = 10^{12}$	TB	$1024^4 = 2^{40}$	TiB (tebibyte)	2^{40}
PB (**petabyte**)	$1000^5 = 10^{15}$	PB	$1024^5 = 2^{50}$	PiB (pebibyte)	2^{50}
EB (exabyte)	$1000^6 = 10^{18}$	EB	$1024^6 = 2^{60}$	EiB (exbibyte)	2^{60}
ZB (zettabyte)	$1000^7 = 10^{21}$	ZB	$1024^7 = 2^{70}$	ZiB (zebibyte)	2^{70}
YB (yottabyte)	$1000^8 = 10^{24}$	YB	$1024^8 = 2^{80}$	YiB (yobibyte)	2^{80}

Test Your Understanding

- Why is the empty text box special, from a software application point of view?

Comparison to State-of-the-art

(from Jan Pedersen’s lecture)



Web search results 1 - 10 of 308 results most relevant to **+darter +habitat**

Next 10 > | Hide summaries | Sort by date | Ungroup results

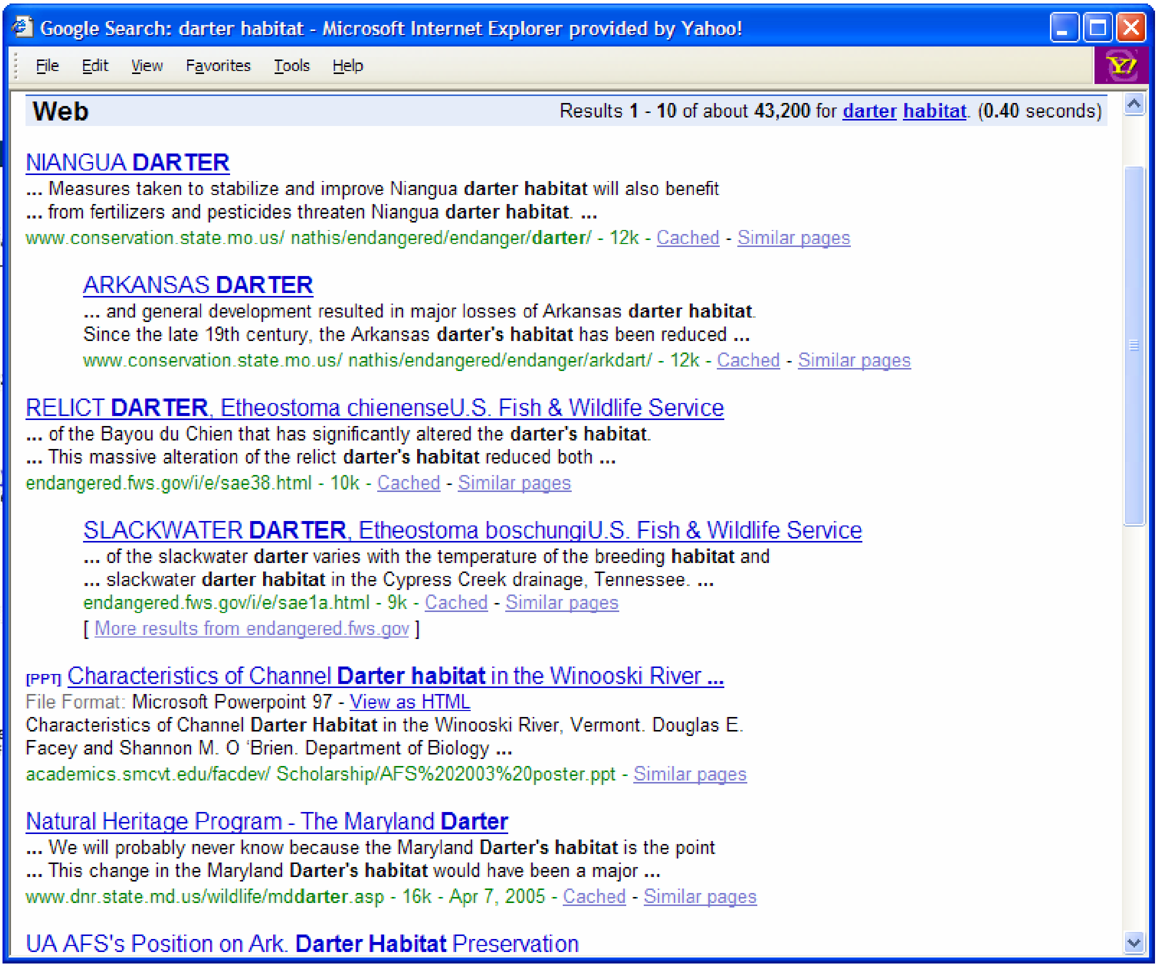
Microhabitat Use in a Diverse Assemblage of Darters ...
Microhabitat Use In A Diverse Assemblage Of Darters In The Elk River Drainage, West Virgini
A. Welsh, Ph.D. Sue A. Perry Rita Villella 1996 - 1997 May 1997 National Biological Service, L
Science Center; W.V. Division of Natural Resources 1. Quantify microhabitat use for ...
92% **Date: 9 Jan 1998**, Size 3.9K, http://www.caf.wvu.edu/coop/elkneck.html
Find similar pages | Grouped results from www.caf.wvu.edu

BAYOU DARTER. Etheostoma (Nothonotus) rubrum U.S. Fish & Wildlife Service
Source: FWS Region 4 -- As of 2/91 Percidae Threatened throughout its range, , September 2
A diminutive species, the Bayou **darter** reaches a maximum length of about 1.8 ...
91% **Date: 13 Apr 1998**, Size 5.1K, http://www.fws.gov/r9endspp/i/e/sae13.html
Find similar pages | Grouped results from www.fws.gov

CWT White-faced Darter Biodiversity Action Plan (Leucorrhinia dubia (Van der Linden))
Cheshire Wildlife Trust White-faced **Darter** Biodiversity Action Plan (Leucorrhinia dubia (Van d
Web Pages 1997.
84% **Date: 5 Jan 1998**, Size 5.5K, http://www.talk-101.com/users/cwt/WfDDBAP.htm
Find similar pages | Grouped results from www.talk-101.com

MDA Pesticide Information Sheet
Address Maryland Department of Agriculture Pesticide Regulation Section 50 Harry S. Truman
Telephone: (410) 841-5710 Fax: (410) 841-2765 Send E-mail to Dennis Howard ...
83% **Date: 15 Nov 1996**, Size 6.1K, http://www.mda.state.md.us/plant/species.htm
Find similar pages

WRCF - Sand Darter
Photo Credit: Rob Criswell IDENTIFYING CHARACTERISTICS: The sand **darter** is a small me
averaging 2 1/2 inches in length. Adults are pale yellow above and silvery below, with a row of



Google Search: darter habitat - Microsoft Internet Explorer provided by Yahoo!
File Edit View Favorites Tools Help

Web Results **1** - **10** of about **43,200** for **darter habitat**. (**0.40** seconds)

NIANGUA **DARTER**
... Measures taken to stabilize and improve Niangua **darter habitat** will also benefit
... from fertilizers and pesticides threaten Niangua **darter habitat**. ...
www.conservation.state.mo.us/ nathis/endangered/endanger/**darter**/ - 12k - Cached - Similar pages

ARKANSAS **DARTER**
... and general development resulted in major losses of Arkansas **darter habitat**.
Since the late 19th century, the Arkansas **darter's habitat** has been reduced ...
www.conservation.state.mo.us/ nathis/endangered/endanger/arkdart/ - 12k - Cached - Similar pages

RELICT **DARTER**, Etheostoma chienenseU.S. Fish & Wildlife Service
... of the Bayou du Chien that has significantly altered the **darter's habitat**.
... This massive alteration of the relict **darter's habitat** reduced both ...
endangered.fws.gov/i/e/sae38.html - 10k - Cached - Similar pages

SLACKWATER **DARTER**, Etheostoma boschungiU.S. Fish & Wildlife Service
... of the slackwater **darter** varies with the temperature of the breeding **habitat** and
... slackwater **darter habitat** in the Cypress Creek drainage, Tennessee. ...
endangered.fws.gov/i/e/sae1a.html - 9k - Cached - Similar pages
[More results from endangered.fws.gov]

[PPT] Characteristics of Channel **Darter habitat** in the Winooski River ...
File Format: Microsoft Powerpoint 97 - View as HTML
Characteristics of Channel **Darter Habitat** in the Winooski River, Vermont. Douglas E.
Facey and Shannon M. O ‘Brien. Department of Biology ...
academics.smcvt.edu/facdev/ Scholarship/AFS%202003%20poster.ppt - Similar pages

Natural Heritage Program - The Maryland **Darter**
... We will probably never know because the Maryland **Darter's habitat** is the point
... This change in the Maryland **Darter's habitat** would have been a major ...
www.dnr.state.md.us/wildlife/md**darter**.asp - 16k - Apr 7, 2005 - Cached - Similar pages

UA AFS's Position on Ark. **Darter Habitat** Preservation

Test Your Understanding

- Why is the search results page unchanged from 10 years ago? Why is it so plain?

Test Your Understanding

- What is needed for high-quality search results?

Test Your Understanding

- What is needed for high-quality search results?
- Good results for:
  - Ranking
  - Comprehensiveness
  - Freshness
  - Presentation

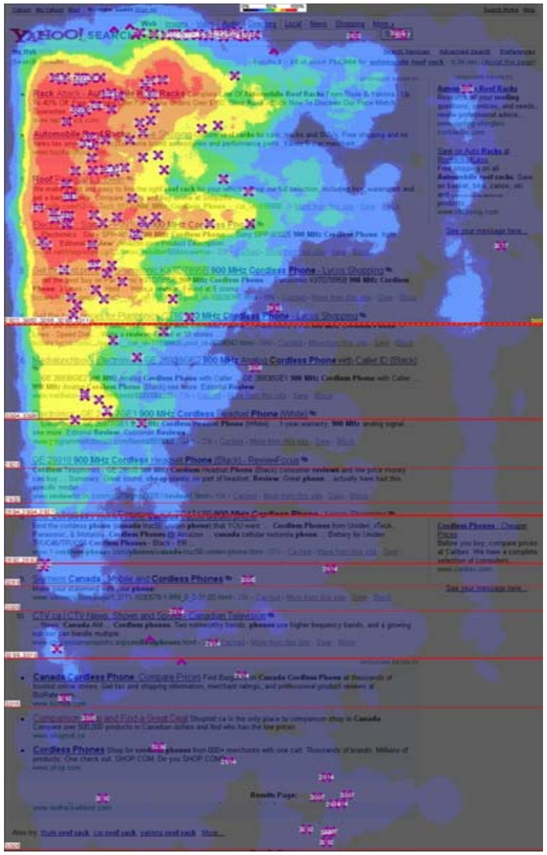
Test your Understanding

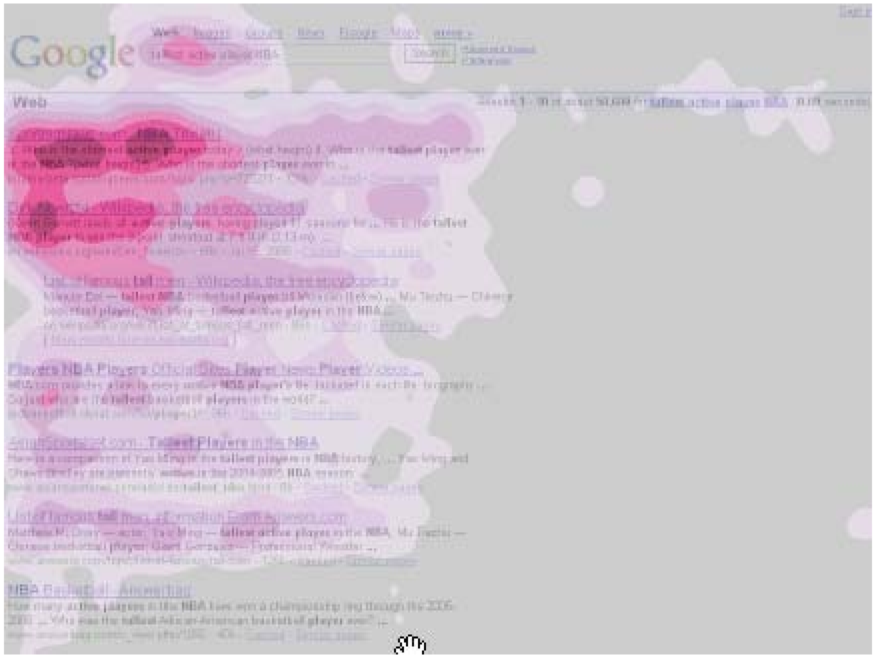
- What are three levels of user evaluation?

Test your Understanding

- What are three levels of user evaluation?
- Micro
  - Small details about the UI; eye tracking
  - Milliseconds
- Meso
  - Field studies
  - Days to weeks
- Macro
  - Millions of users
  - Days to months

What do these mean?





Test your understanding

- What is meant by ambiguous and disambiguate?

Test your understanding

- What is meant by ambiguous and disambiguate?
  - Words with more than one meaning or more than one sense
    - Jets: sports team or airplane?
    - Bass: fish or musical instrument?

Test Your Understanding

- What is morphological analysis, also known as stemming?

Test Your Understanding

- What is morphological analysis, also known as stemming?
  - Convert a word to its base form:
    - Running, ran, runs -> run
    - Building, builder, builds -> build? Not always

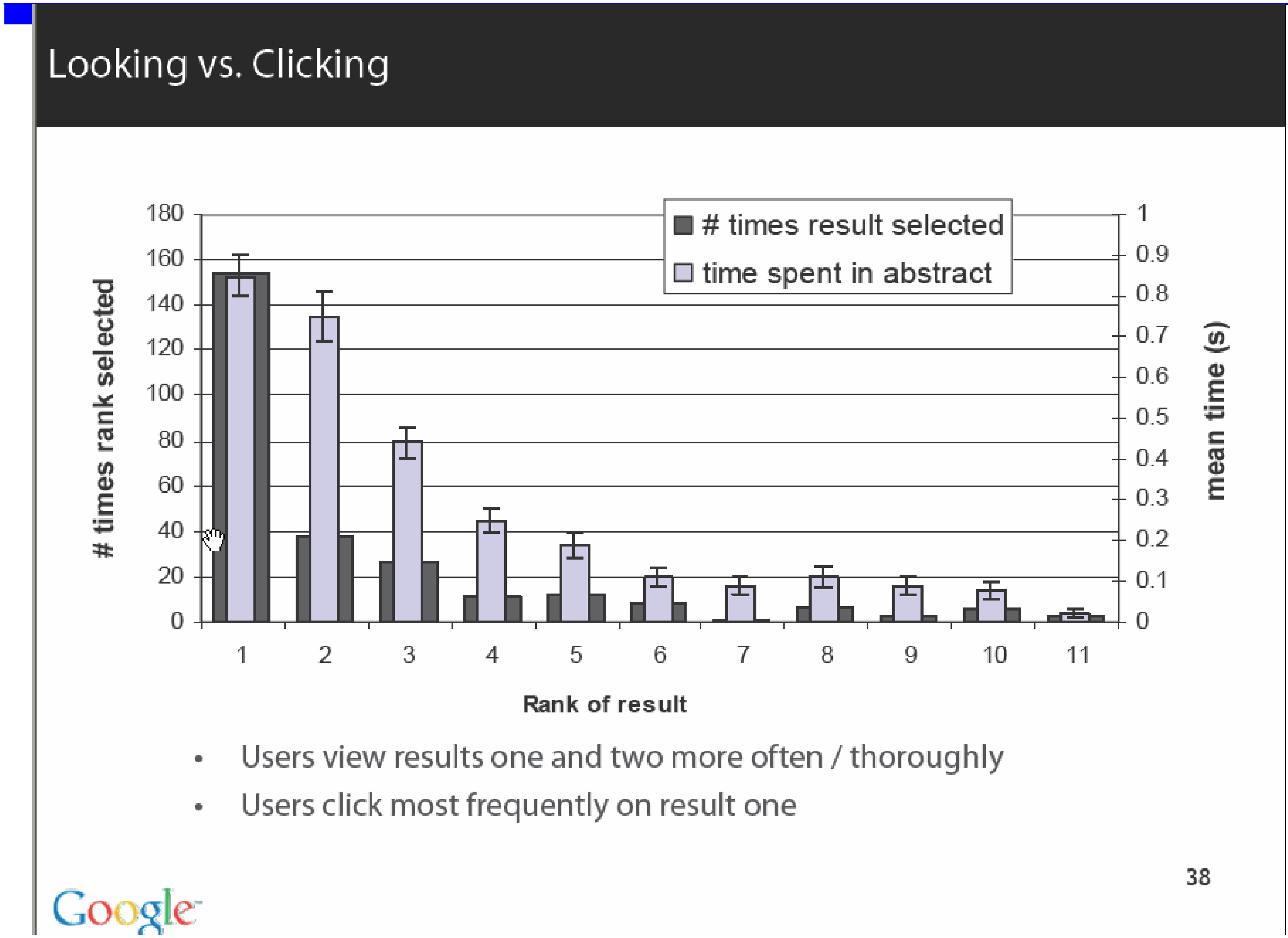
Test Your Understanding

- Why is it not always a good idea to stem query terms?

Test Your Understanding

- Why is it not always a good idea to stem query terms?
  - Sometimes the form a word is used int indicates something about the sense of the word.
    - apple vs apples

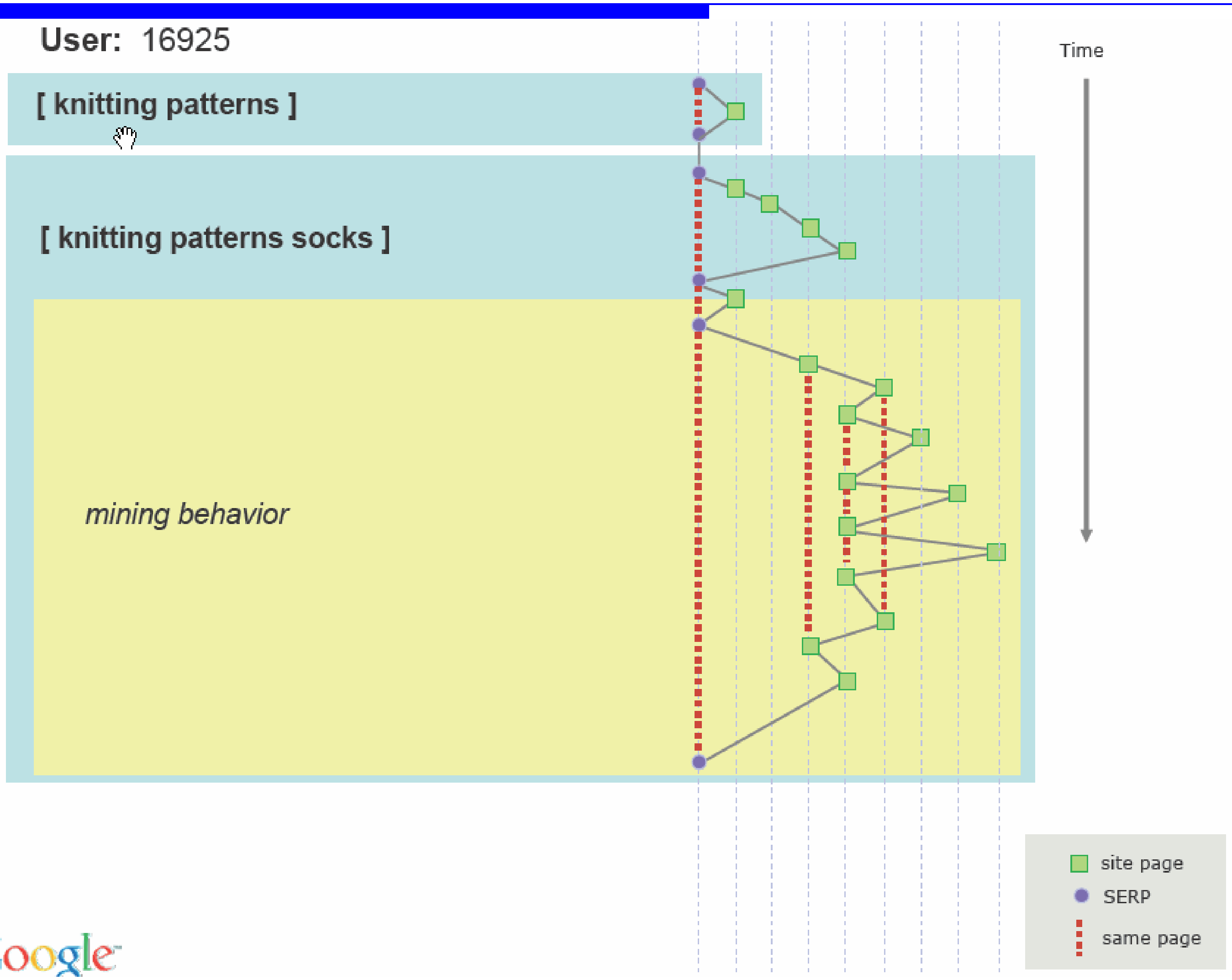
What does this mean? (from Dan Russell’s lecture)



Looking vs. Clicking

- Users view results one and two more often / thoroughly
- Users click most frequently on result one

What does this mean? (from Dan Russell's lecture)



Test Your Understanding

- Are peoples' search skills evolving? If so, how?
- What is "teleporting?" What is "orienteering?"

Test Your Understanding

- What are navigational queries?

- What other kinds of queries are there?

- What do these queries mean?
  - banana
  - Sgt Peppers Lonely Hearts Club Band
  - Why is my dog sick?

Search Operators

- How do “double quotes” work?
- What does * mean?
- What is AND vs. OR?

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

Slide adapted from Lew & Davis