CSE 454
Advanced Internet & Web Services

- **Prof: Dan Weld**
  - Most lectures, concepts, perspective.
- **TA: Alan Liu**
  - Machine/environment/software, project details
- **Expectations:**
  - Project (multiple parts, on time!)
  - Reading (papers, web - no formal text)
  - Class participation / development
- **Caveat: Life on the cutting edge**

My Background

- **Research on Intelligent Internet Systems [1991-**
  - Internet Softbot (Discover award finalist '95)
  - Webcrawler by Brian Pinkerton
  - Metacrawler by Eric Selberg & Oren Etzioni
  - Mulder (first automated WWW question answerer)
  - KnowItAll - massive, autonomous information extraction
- **Co-founded**
  - Netbot
  - AdRelevance
  - Nimble Technology
  - Asta Networks
- **Leaves of absence**
  - VP Engineering at Netbot
  - Venture Partner w/ Madrona Venture Group.
- **Incredible shortage of software engineers!**
- **Dearth of training**

Your Background?

- **Classes?**
  - 444, 451, 461, 473
- **Concepts?**
  - Threads, race condition, deadlock
  - Naïve Bayes classifier
  - Hybrid hash join algorithm
  - Precision, recall
  - Fingerprint algorithm
  - LRU cache replacement policy
- **Programming Background?**
  - Java, .NET, J2EE, XML, admin own webserver

Course Outcomes

- **After this course, you should know:**
  - How search engines work
  - How to build scalable web sites
  - How Amazon generates personalized recommendations
  - How digital cash works
  - Issues in e-commerce
  - How to build peer2peer systems (overlay networks)
- **Focus: search! (why?)**

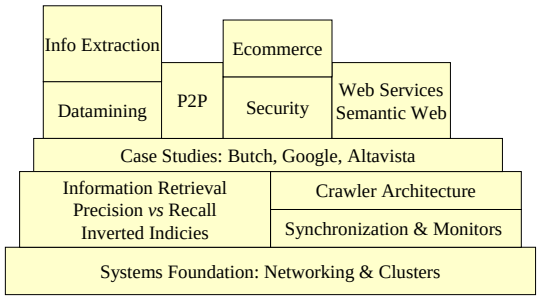
Why Search?

- **A billion or so searches per day…**
- **Boost to productivity**
  - Intellectual & economic
- **Search is 'hot'**
  - Google, Amazon, Ebay,
  - Search for/in books, products, music, people, …
- **Fascinating research problem.**
  - Research complete: systems + AI
- **You can learn to be a something of a search expert in one quarter!**

Syllabus

- **Introduction**
  - History, networking overview, web server architecture
- **Information Retrieval on the Web**
  - Crawling, indexing, scaleup issues
  - Vector space model,
  - Hyperlink analysis
- **Data Mining**
  - collaborative filtering, clustering, classification
- **Web Services**
  - Protocols, brokers, meta-search, data integration
- **Information Extraction**
  - Question answering
  - The future of search
- **Special Topics (Time permitting)**
  - Semantic web, e-commerce, security, peer-to-peer, advertizing

Course Overview



What This Course Is Not

… there is a difference between training and education. If computer science is a fundamental discipline, then university education in this field should emphasize enduring fundamental principles rather than transient current technology.
-Peter Wegner, Three Computing Cultures. 1970.

- **We won't:**
  - Teach you how to be a web master
  - Teach all the latest x-buzzwords in technology
    - XML/SOAP/WSDL
      - (okay, may be a little).
  - Teach web/javascript/java/jdbc… programming

Grading

- **Group Project**
  - 85% Project (Homeworks)
    - Part artifact
    - Part writeup
      - Clear and concise explanation / justification
      - Experimentation
  - 15% Class participation
- **Note: 454 is a capstone design class**

Default (Group) Project

- **Mini Google**

1. **Create Inverted Index**
2. **Ranking: IR++, Hyperlink analysis**
3. **Search Mining: apply ML to … ?**
   - Text catigorization ?
   - Clustering search results ?
   - Information extraction ?
   - ????

Or…. Do your own thing

- **Search UW library**
- **Search MSFT Help**
- **Search for Webcams**
- **?????**

- **But:**
  - Move fast
  - Write one-page proposal, due in 1 week
  - Milestones are crucial

Warning

- **No textbook**
- **Large project component**
- **Poorly documented, unstable systems**
- **Field changes quickly**
  - Each year is essentially a new course
- **Need students to help debug class!**

Ancient History

Pre-history: Census, Dewey Decimal system
and other bizarre medieval rituals performed by hand.
1950s: "Information Retrieval" (IR) term coined
1960 Ted Nelson proposes Xanadu
Hypertext vision of WWW
1961 Kleinrock paper on packet switching
Contrast with phone lines, which are circuit switched.
1965 Gordon Moore proposes law
1966 Design of ARPAnet

History 2

1968 Doug Engelbart: the first WIMP
Gerald Salton SMART system (Cornell)
vector space model, "father of IR"
1969 First ARPAnet message UCLA -> SRI
1970 ARPAnet spans country, has 5 nodes
1971 ARPAnet has 15 nodes
1972 First email programs, FTP spec
1973 Ethernet operation at Xerox PARC

History 3

1974 Intel launches 8080;
TCP design
1975 Gates/Allen write Basic for Altair 8800
1976 Apple Computer formed by Jobs/Wozniak
1977 111 hosts on ARPAnet
1979 Visicalc
1980s: Proprietary document DBs
Lexis-Nexis, Medline
1981 Microsoft has 40 employees;
IBM PC

History 4

1983 ARPAnet uses TCP/IP
Birth of internet
1983 Design of DNS
1984 Launch of Macintosh;
1000 hosts on ARPAnet
1985 Symbolic.com first registered domain name
1989 100,000 hosts on Internet
1990 Cisco Systems goes public $288 M
Tim Berners-Lee creates WWW at CERN
Archie (index file names, anon. ftp servers)

History 5

1991: Gopher (menus, links, to servers)
1992: Veronica (index of menu items on gophers)
1993: Jughead (keyword + boolean search)
Mosaic browser developed at UIUC
Web grows by 341,000% in a year
WWW Wonderer (first crawler)
1994 Webcrawler built (UW class project!)
Yahoo (directory) launched,
Netscape & Amazon formed
1995 Netscape IPO,
Windows 95,
Ebay founded
MetaCrawler built (UW MS thesis)

Recent History

1997: Goto.com ("sponsored links" pay-per-click)
AskJeeves (question answering)
Netbot (comparison-shopping search)
Amazon IPO
1998: Open directory launched
Google, pagerank algorithm
1999: SE becomes portal (Yahoo, Excite)
"Search is a commodity"
2000: Flipdog (information extraction)
2001: Ascendance of Google
"search is nirvana"
Dominance of advertising model

Approaching the Present

2002+:

Image Search
Dating sites (person search)
Peer-peer systems
VoIP
Web Services
Local search
Browsing on mobile devices (cellphone, etc)
Link-Spamming (Arms race to bias SE ranking)
Social Networking Sites
Desktop search
Search for Maps
Tagging
Digital Earth

The Future?

Video Google

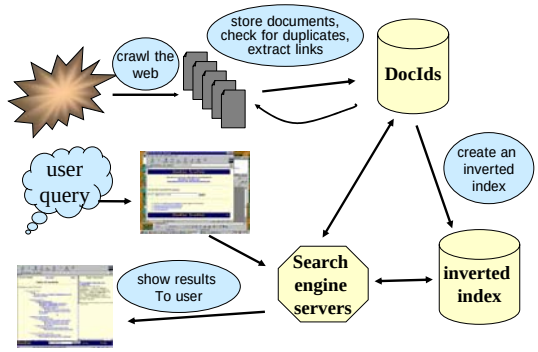
http://www.robots.ox.ac.uk/~az/talks/sicily.html

???

Search Engine Overview

- **Spider**
  - Crawls the web to find pages. Follows hyperlinks. Never stops
- **Indexer**
  - Produces data structures for fast searching of all words in the pages
- **Retriever**
  - Query interface
  - Database lookup to find hits
    - 300 million documents
    - 300 GB RAM, terabytes of disk
  - Ranking, summaries
- **Front End**

Standard Web Search Engine Architecture



Slide adapted from Marty Hearst / UC Berkeley]

Spiders (Crawlers, Bots)

- **Queue := initial page URL_0**
- **Do forever**
  - Dequeue URL
  - Fetch P
  - Parse P for more URLs; add them to queue
  - Pass P to (specialized?) indexing program

- **Issues**
  - Which page to look at next? (keywords, recency, ?)
  - Avoid overloading a site
  - How deep within a site to go (drill-down)?
  - How frequently to visit pages?
  - Traps!

Retrieval (Conceptually)

- Document-term matrix

	t_1	t_2	...	t_j	...	t_m	nf
d_1	w_{11}	w_{12}	...	w_{1j}	...	w_{1m}	$1/\|d_1\|$
d_2	w_{21}	w_{22}	...	w_{2j}	...	w_{2m}	$1/\|d_2\|$
	. . . . .						
d_i	w_{i1}	w_{i2}	...	w_{ij}	...	w_{im}	$1/\|d_i\|$
	. . . . .						
d_n	w_{n1}	w_{n2}	...	w_{nj}	...	w_{nm}	$1/\|d_n\|$

- w_{ij} is the weight of term t_j in document d_i
- Most w_{ij}'s will be zero.

Inverted Index

FILE

POS	
1	A file is a list of words by position
10	First entry is the word in position 1 (first word)
20	Entry 4562 is the word in position 4562 (4562nd word)
30	Last entry is the last word
36	An inverted file is a list of positions by word!

a (1, 4, 40)
entry (11, 20, 31)
file (2, 38)
list (5, 41)
position (9, 16, 26)
positions (44)
word (14, 19, 24, 29, 35, 45)
words (7)
4562 (21, 27)

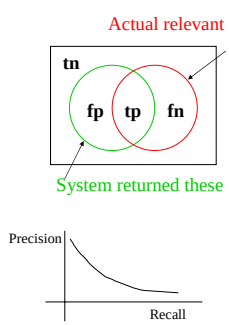
INVERTED FILE

Ranking models in IR

- **Key idea:**
  - *We wish to return in order the documents most likely to be useful to the searcher*
- **To do this, we want to know which documents *best* satisfy a query**
  - An obvious idea is that if a document talks about a topic *more* then it is a better match
- **A query should then just specify terms that are relevant to the information need, without requiring that all of them must be present**

Precision & Recall

- **Precision** $\frac{tp}{tp + fp}$
  - Proportion of selected items that are correct
- **Recall** $\frac{tp}{tp + fn}$
  - Proportion of target items that were selected
- **Precision-Recall curve**
  - Shows tradeoff



Precision-recall curves

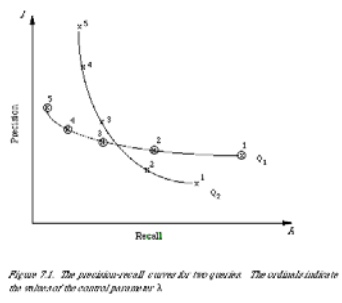


Figure 7.1. The precision-recall curves for two queries. The coordinate indicates the influence of the control parameter λ

Ranking

- **Term Frequency**
- **Text on the page *vs…***
- **Link structure**
- **???**

For next time

- **Add yourself to mailing list**
- **Think about project**
- **Think about groups**