

CSE 454 Advanced Internet & Web Services



CSE 454 Advanced Internet & Web Services



CSE 454 Advanced Internet & Web Services

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

4/1/2009 9:48 AM

3

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
 - Semantic Wikipedia
- **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**

4/1/2009 9:48 AM

4

Your Background?

- **Classes?**
 - 444, 446, 451, 461, 473, 490H
- **Concepts?**
 - Threads, race condition, deadlock
 - Naïve Bayes classifier
 - Hybrid hash join algorithm
 - Precision, recall
- **Programming Background?**
 - Ruby, .NET, XML, admin own webserver

4/1/2009 9:48 AM

5

Topics

3/31	Introduction: overview, mechanics + history	4/2	Introduction to IR: precision/recall, cross validation, cosine, inv index
4/7	Maxim: Machine Learning, Naive Bayes, Text Categorization	4/9	Jessa Davis: Issues in ML: overfitting, ensembles, cotraining, clustering (k means, STC)
4/14	Information Extraction: Overview, sliding window & rule learning	4/16	Information Extraction with FS models
4/21	No Class: Group Meetings	4/23	Chloe Kiddon: POS Tagging & Parsing
4/28	Open IE: Bootstrapping & Self-supervision	4/30	Mike Cafarella: Webtables & Octopus
5/5	Mike Mathieu: Online Advertising	5/7	No Class: Group Meetings
5/12	Web: HTTP, servers, scaling, crawling	5/14	Search Engines: Indexing
5/19	Search engines: Case studies (Google & Altavista), Link analysis	5/21	Josh Benaloh: Cryptography & Security
5/26	No Class: Group Meetings	5/28	Jaime Teevan: Revisitation & Personalized Search
6/2	Eytan Adar: Zoetrope	6/4	Summary & Bonus Topic: Human Computation

Key Topics

- Machine Learning: 12%
- Information Extraction: 24%
- IR & Web Search: 29%

Course Outcomes

- After this course, you should know:
 - How search engines work
 - How to build information extraction systems
 - How to ensure a web site scales
 - How Amazon generates personalized recommendations
 - Cryptography fundamentals
 - Other cool stuff
- Focus: search! (why?)

4/1/2009 9:48 AM

8

Why Search?

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

4/1/2009 9:48 AM

9

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

NAME	TITLE	ORGANIZATION
------	-------	--------------

Slides from Cohen & McCallum

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Slides from Cohen & McCallum

Why Information Extraction

- Next-Generation Search
 - People
 - Zoominfo
 - Flipdog
 - Intellius
 - Research Papers
 - Citeseer
 - Google scholar
 - Product search
- Question Answering

4/1/2009 9:48 AM

12

Example

The screenshot shows the ZoomInfo website interface. At the top, there are navigation tabs for 'Company Search', 'People Search', and 'Job Search'. Below this, a search bar contains the name 'Daniel weld'. The main content area displays a list of results for 'Daniel Weld', with 14 of 16 people shown. Each result includes a name, title, and company. For example, 'Weld, Daniel S.' is listed as a 'Venture Partner' at 'Madrona Venture Group LLC'. There are also filters on the left side for 'Geography' and 'Annual Revenue'.

4/1/2009 9:48 AM

13

...Continued

This screenshot continues the ZoomInfo search results for Daniel S. Weld. It shows a detailed profile for 'Dr. Daniel S. Weld', including his title as 'Venture Partner' at 'Madrona Venture Group LLC'. Below the profile, there are sections for 'Employment History', 'References', and 'Member, Computer Science and Engineering Department'. The 'References' section lists several publications and articles, such as 'A Scalable Based Interface to the Internet' and 'An Approach to Planning with Incomplete Information'. The 'Member' section lists various professional affiliations, including 'Member of the Faculty of Computer Science and Engineering' at the University of Washington and 'Chief Scientist' at 'Adlevance Inc.'.

4/1/2009 9:48 AM

...Continued Some More

This screenshot continues the ZoomInfo search results for Daniel S. Weld. It shows a detailed profile for 'Dr. Daniel S. Weld', including his title as 'Venture Partner' at 'Madrona Venture Group LLC'. Below the profile, there are sections for 'Employment History', 'References', and 'Member, Computer Science and Engineering Department'. The 'References' section lists several publications and articles, such as 'A Scalable Based Interface to the Internet' and 'An Approach to Planning with Incomplete Information'. The 'Member' section lists various professional affiliations, including 'Member of the Faculty of Computer Science and Engineering' at the University of Washington and 'Chief Scientist' at 'Adlevance Inc.'.

4/1/2009 9:48 AM

15

CiteSeer vs. Scholar

This screenshot compares search results from CiteSeer and Google Scholar. The top part shows CiteSeer search results for 'Daniel Weld', listing several papers such as 'A Scalable Based Interface to the Internet' and 'An Approach to Planning with Incomplete Information'. The bottom part shows Google Scholar search results for the same query, displaying a list of articles with their titles, authors, and publication dates. The comparison highlights the differences in the way these two search engines present and index academic literature.

Grading

- 85% Project (Staged in Parts)
 - Part artifact
 - Part writeup
 - Clear and concise explanation / justification
 - Experimentation
 - Part presentation
- 15% Class participation

4/1/2009 9:48 AM

17

Capstone Projects

- Done in Group
 - Why?
- Default Topics
 - Information Extraction Orientation
 - But you can roll your own - see me
- Hadoop
 - Optional, but we have the cluster

4/1/2009 9:48 AM

18

Start with Concrete Problem

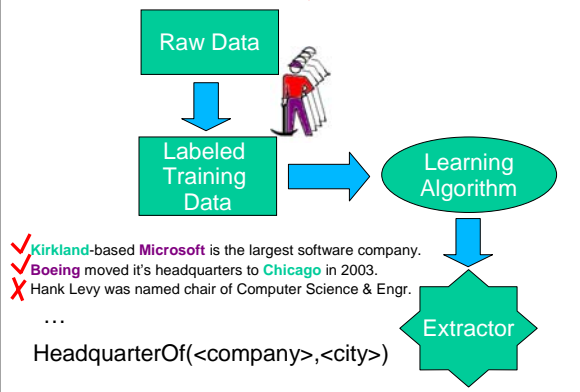
- Text Classification
- Corpus of Wikipedia pages
 - E.g., scientist, writer, author, university
- You'll use machine learning to construct
 - Program which outputs the 'type' of the page
- Details later this week.

Project Possibilities

- Extract Facts from Wikipedia
 - Or recipes, or ...?
- Build Ontology of Products & Attributes
- Mine product reviews for attribute valence
- Recommend Twitter feeds
- Or suggest something different

Teams & ideas settled by 4/14

Traditional, Supervised I.E.



Kylin: Self-Supervised Information Extraction from Wikipedia

[Wu & Weld CIKM 2007]



From infoboxes to a training set

Clearfield County, Pennsylvania	
Statistics	
Founded	March 26, 1804
Seat	Clearfield
Area	
- Total	2,988 km ² (1,154 mi ²)
- Land	sq mi (km ²)
- Water	17 km ² (6 mi ²), 0.56%
Population	
- (2000)	83,382
- Density	28/km ²

Clearfield County was created in 1804 from parts of Huntingdon and Lycoming Counties but was administered as part of Centre County until 1812.

Its county seat is Clearfield.

2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water.

As of 2005, the population density was 28.2/km².

New York City hotels > Mandarin Oriental New York

Opine

Review Summary

Service quality: [excellent \(3\)](#), [good \(2\)](#), [best](#), [professional](#), [better](#), [view all](#)

Service attention: [attentive \(2\)](#)

Room beauty: [absolutely beautiful](#), [beautiful](#), [view all \(2\)](#)

User comments:

The service was excellent and our room was absolutely beautiful. [Read more](#)

When compared to Mandarin Oriental New York, Room beauty is

- worse at The Premier (33 others)

Quality: [best](#), [finest](#), [love](#), [better](#), [view all \(4\)](#)

Staff courtesy: [extremely courteous](#), [courteous](#), [view all \(2\)](#)

Beauty: [beautiful](#)

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

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!

4/1/2009 9:48 AM

25

Ancient History

- Pre-history: Dewey Decimal system
 - and other bizarre medieval rituals performed by hand
- 1960: Ted Nelson proposes Xanadu
 - Hyperext vision of WWW---why did it fail?
 - Focus on copyright issues (still a thorny problem)
 - Focus on stable, bidirectional links
 - "Trying to fix HTML is like trying to graft arms and legs onto hamburger"-- Ted Nelson

1961 Kleinrock paper on packet switching

Contrast with phone lines - circuit switched.

4/1/2009 9:48 AM

26

Paleolithic Era

- 1965 Gordon Moore proposes law
- 1966 Design of ARPAnet
- 1968 Doug Engelbart: the first WIMP
- 1969 First ARPAnet message UCLA -> SRI
- 1970 ARPAnet spans country, has 5 nodes
- 1971 ARPAnet has 15 nodes
- 1972 First email programs, FTP spec

4/1/2009 9:48 AM

27

The Personal Computer Era

- 1974 Intel launches 8080;
 - TCP design
- 1975 Gates/Allen write Basic - Altair 8800
- 1976 Jobs/Wozniak form Apple Computer
 - 111 hosts on ARPAnet
- 1979 Visicalc
- 1981 Microsoft has 40 employees;
 - IBM PC
- 1984 Launch of Macintosh
- 1986 Microsoft goes public

4/1/2009 9:48 AM

28

Internet ramps up

- 1983 ARPAnet uses TCP/IP, Design of DNS
 - 1000 hosts on ARPAnet
- 1985 Symbolic.com first registered domain name
- 1989 100,000 hosts on Internet
- 1990 Cisco Systems goes public
 - Tim Berners-Lee creates WWW at CERN

4/1/2009 9:48 AM

29

Web Search Pre-History

- 1950s: "Information Retrieval" (IR) term coined
- 1960s-70s: SMART system, vector space model,
 - Gerald Salton (Cornell) father of IR
- 1980s: Proprietary document DBs
 - (Lexis-Nexis, Medline)
- 1990: Archie (index file names, anon. ftp)
- 1991: Gopher (menus, links to servers)
- 1992: Veronica (index of menu items on gophers)
- 1993: Jughead (keyword + boolean search)
- Rapid evolution, but what is missing?

4/1/2009 9:48 AM

30

Modern History of Search

- 1993: WWW Wanderer (first crawler)
- 1994: WebCrawler, Lycos (1st widely-used SEs)
 - WebCrawler was a UW class project by Brian Pinkerton
- 1994: Yahoo directory (Stanford; founded '95)
Amazon founded
Netscape founded (90% mkt share → 1%)
- 1995: Ebay
MetaCrawler (1st major meta-SE)
 - UW Master's thesis by Erik Selberg

4/1/2009 9:48 AM

31

Discovery of the Biz Model

- 1996: Flash by Macromedia
 - later acquired by Adobe
- 1997: goto.com
 - "sponsored links" pay-per-click
- 1997: AskJeeves (question answering)
- 1997: Netbot
 - comparison-shopping search
- 1998: Open directory launched
Google, pagerank algorithm
Paypal founded

Turn of the Millennium

- 1999: IE becomes dominant browser
Napster starts operation
Search Engines → portals (Yahoo, Excite)
"Search is a commodity"
- 2000: Flipdog
 - commercial information extraction)
- 2001: Bittorrent protocol (now 35% of internet)
Ascendance of Google
"Search is nirvana"
- 2002: IE peaks at 90% market share



4/1/2009 9:48 AM

33

Approaching the Present

- 2003: Skype released
- 2004: Facebook founded
Social news (Digg)
- 2005: Youtube founded
 - 9.5 B videos shown per month
 - 33 months after founding!
- 2006: Twitter founded
- 2007: Google Streetview
Apple iPhone
- 2009: Facebook 200M users



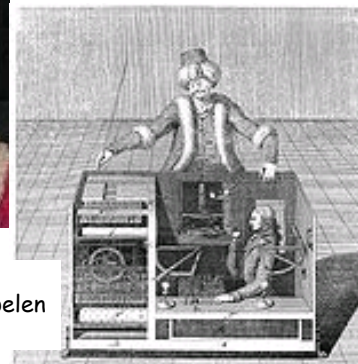
Future of the Net

- Domination of Mobile Devices (cellphone, etc)
- Link-Spamming (Arms race to bias SE ranking)
- Local Search, Digital Earth
- Image & Video search
- Social news (Digg / Twitter)
- Crowd Sourcing
- What else?

4/1/2009 9:48 AM

35

Mechanical Turk



Built in 1770 by
Wolfgang von Kempelen

4/1/2009 9:48 AM

- **Launched in Nov '05**
 - Initially: detect duplicate product pages
- **100k workers in 100 countries by 3/07**
 - 34k HITs on 3/28/08
- **Search for Jim Gray**
 - 12k searchers

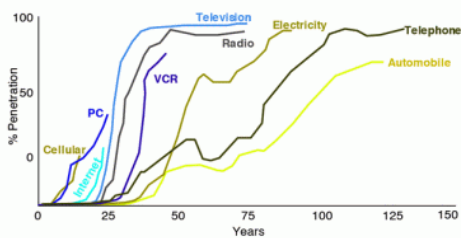
Observations

- **Internet/Web evolved** - it wasn't created
- **Scalability beats structure**
 - search engines over directories
 - Web over hypertext
- **"We are 10 seconds from the Big Bang"**
 - John Doerr

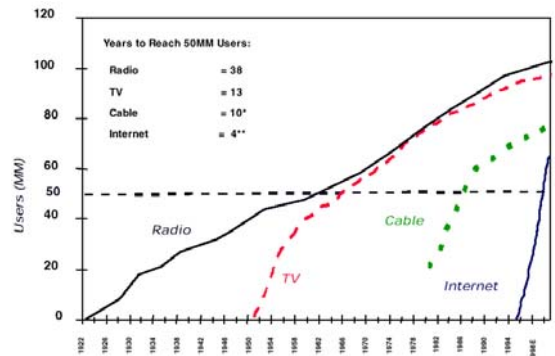
Adoption

Facilitating Innovation the pace of innovation is increasing

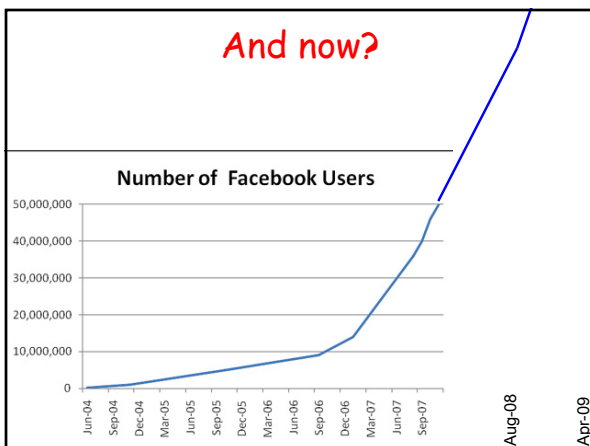
- Newer technologies taking hold at double or triple previous rates



Accelerating

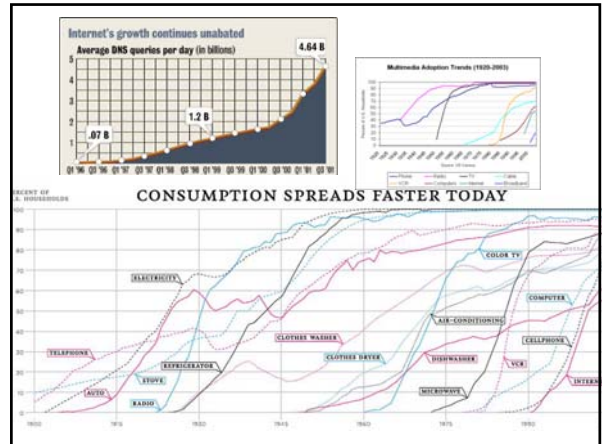
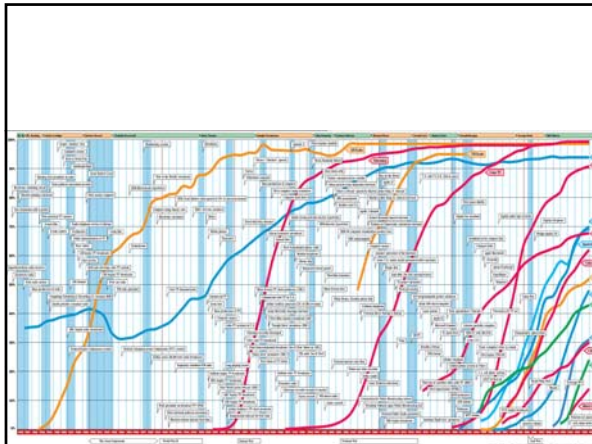


And now?



For Next Time

- **Add yourself to mailing list**
 - We'll send out a key email tomorrow
 - Be sure to get it
- **Think about project**
 - Form a group (3-4 people)



33 months after founding

Top U.S. Online Video Properties* by Videos Viewed
November 2007
Total U.S. - Home / Work / University Locations
Source: comScore Video Metrix

Property	Videos Viewed (MM)	Share (%) of Videos
Total Internet	9,491	100.0%
Google Sites	2,966	31.3%
Fox Interactive Media	419	4.4%
Yahoo! Sites	328	3.5%
Viacom Digital	245	2.6%
Time Warner Network	184	1.9%
Microsoft Sites	181	1.9%
Disney Online	96	1.0%
ABC.com	88	0.9%
ESPN	87	0.9%
Break	47	0.5%