

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: Jessica Leung**
 - Project details
- **Expectations:**
 - Project (multiple parts, *on time!*)
 - Reading (papers, web - no formal text)
 - Class participation / development
- **Caveat: Life on the cutting edge**

10/1/2009 4:59 PM

3

My Background

- **Research on Intelligent Internet Systems [1991-**
 - Internet Softbot
 - Discover Award Finalist '95
 - Webcrawler
 - By Brian Pinkerton
 - Metacrawler & Shopbot
 - Basis for Netbot Inc.
 - Mulder
 - First automated WWW question answerer
 - KnowItAll
 - Massive, autonomous information extraction
 - Intelligence in Wikipedia Project



10/1/2009 4:59 PM

4

Background Continued

- **Co-founded**
 - Netbot (Jango)
 - 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**



(r)



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

10/1/2009 4:59 PM

6

454 Topics

- Information Retrieval
- Search Engines
 - Crawling, Indexing, Query Processing, Ranking
 - Pagerank, Interfaces
- Text Categorization & Clustering
- Information Extraction
 - Machine Learning
- Internet Advertising
- Security, Cryptography, Malware
- Social Networks
- Temporal Web
- Special Topics

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?)

10/1/2009 4:59 PM

8

Why Search?

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

10/1/2009 4:59 PM

9

What is "Information Extraction"

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

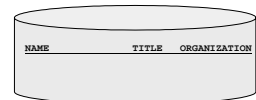
October 14, 2002, 4:00 a.m. PT

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...



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
 - Intelius
 - Research Papers
 - Citeseer
 - Google scholar
 - Product search
- Question Answering

10/1/2009 4:59 PM

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 is a search bar with 'Person Name: Daniel weld' and a search button. A filter sidebar on the left allows for refining results by 'Geography' (set to 'United States') and 'Annual Revenue' (set to 'No Min' to 'No Max'). The main results table shows 14 of 16 people, with columns for Name, Title, and Company. The first entry is Daniel S. Weld, Venture Partner at Madrona Venture Group LLC. Other entries include Associate Editor at AI Access Foundation Inc, and various roles at the University of Washington and Northwestern University.

10/1/2009 4:59 PM

13

...Continued

This block continues the profile for Daniel S. Weld. It includes a 'Member, Computer Science and Engineering Department' at the University of Washington, a 'Member of the Faculty of Computer Science and Engineering' at the University of Washington, and a 'Chief Scientist' role at Adlevance Inc. It also lists 'Founding' roles at Helbot Inc and 'Co-Founder' roles at Journal of AI Research. A 'Program Chair' role is mentioned for the American Association for Artificial Intelligence. The 'Board Membership and Affiliations' section lists the Madrona Venture Group LLC. Under 'Member of Technology Advisory Board', it lists Madrona Venture Group LLC. The 'Architecture Committee Member' role is listed at DAML. The 'Fellow (past)' section mentions the American Association for Artificial Intelligence. A bio paragraph states: '1999, an Office of Naval Research Young Investigator's award in 1995, was elected a AAAI Fellow in 1995, and an ACM Fellow in 2006. He earned bachelor's degrees in both Computer Science and Biochemistry at Yale University in 1982, and a Ph.D. from the Massachusetts Institute of Technology's Artificial Intelligence Lab in 1988. Dr. Weld is Co-Founder of the Journal of AI Research and is on the Editorial Board of Artificial Intelligence.' A 'Web' link points to 'View all 47 references'. The 'References' section lists a paper by Daniel S. Weld, 'A Software-Based Interface to the Internet', published in the Journal of AI Research in 1995. The 'Employment History' section lists 'Venture Partner' at Madrona Venture Group LLC from 2006 to present, and 'Associate Editor' at AI Access Foundation Inc from 2005 to present. The 'Education' section lists 'Ph.D.' from the Massachusetts Institute of Technology (1988) and 'Bachelor's degrees, Computer Science and Biochemistry' from Yale University (1982).

10/1/2009 4:59 PM

...Continued Some More

This block continues the profile for Daniel S. Weld. It lists 'Education' with a 'Ph.D.' from the Massachusetts Institute of Technology (1988) and 'Bachelor's degrees, Computer Science and Biochemistry' from Yale University (1982). It also lists '4. Nimble Technology Tech. Advisory Board - provides 30K data integration software for real-time unified views of database, data warehouses, and unstructured sources. Create enterprise information portals, business intelligence and other applications.' and '5. Adlevance - press releases intelligence advance.com - (cached)' published on 11/4/2002. It mentions 'The CMISAC technology was developed by a team of engineers led by Adlevance Chief Scientist, Dan Weld, Ph.D. and Vice President of Engineering Jiw Bant'. A bio paragraph states: 'Weld and Bant are probably best known for their work in bringing Jango, Helbot's intelligent shopping agent, to market in 1997.'

10/1/2009 4:59 PM

15

CiteSeer vs. Scholar

The screenshot shows the CiteSeer search results for 'Daniel Weld'. The search criteria are 'Documents' and 'Citations'. The results list several papers, including 'A Software-Based Interface to the Internet' by Daniel Weld (1995), 'An Approach to Probabilistic Planning' by Kulkarni, Hanks, and Weld (1995), 'Planning in Hybrid Domains' by Steve Hanks, Daniel Weld, and others (1995), 'A Scalable Comparison-Shopping Agent' by Dan Eason, Orit Etzion, and Daniel Weld (1995), 'An Approach to Planning with Incomplete Information' by Eason, Hanks, and Weld (1995), 'An Adaptive Query Execution System for Data Integration' by Eason, Hanks, and Weld (1994), 'An Algorithm for Probabilistic Least-Commitment Planning' by Kulkarni, Hanks, and Weld (1994), and 'An Introduction to Least-Commitment Shopping' by Dan Eason, Orit Etzion, and Daniel Weld (1995). The 'Scholar All articles - Recent articles' section shows a list of results with links to the full text, PDF, or HTML versions.

Grading

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

10/1/2009 4:59 PM

17

Capstone Projects

- Done in Group
 - Why?
- Topics
 - Roll your own
 - Or see me

10/1/2009 4:59 PM

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 online
 - Done in pairs
 - Due 10/13

Project Possibilities

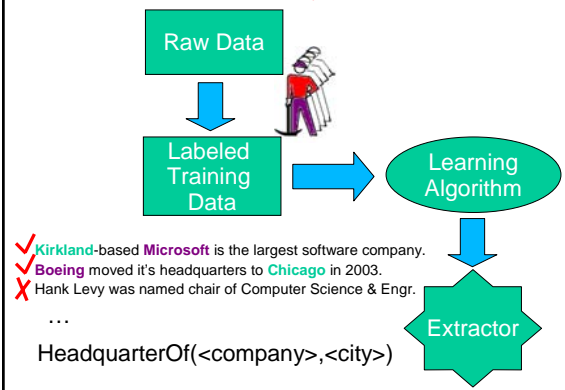
- Extract Facts from Wikipedia
 - Or recipes, or ...?
- Build Ontology of Products & Attributes
- Mine product reviews for attribute valence
- Or suggest something different

Teams & ideas settled by 10/13

Last Quarter's Projects

- Craigslist++
- University Search
- Twitter Feedrank
- Apartment Listing & Aggregation
- Webcam Identification & Search
- Trail / Hike Search
- Seattle Event Finder
- Automatic Stock Investor

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, 1504
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

Review Summary

Opine

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

10/1/2009 4:59 PM

25

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!

10/1/2009 4:59 PM

26

Ancient History

- **Pre-history: Dewey Decimal system**
 - Bizarre medieval rituals performed by hand
- **1960: Ted Nelson → Xanadu**
 - Hypertext 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.

10/1/2009 4:59 PM

27

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



10/1/2009 4:59 PM

28

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

10/1/2009 4:59 PM

29

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

10/1/2009 4:59 PM

30

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?

10/1/2009 4:59 PM

31

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

10/1/2009 4:59 PM

32

Discovery of the Biz Model

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

Turn of the Millennium

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

10/1/2009 4:59 PM

34

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




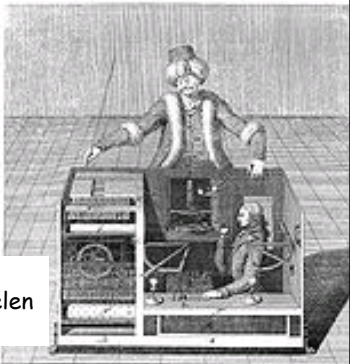
Future of the Net

- 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?

10/1/2009 4:59 PM

36

Mechanical Turk


Built in 1770 by Wolfgang von Kempelen

10/1/2009 4:59 PM

amazon mechanical turk

beta Artificial Intelligence

- 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



10/1/2009 4:59 PM 38

Observations

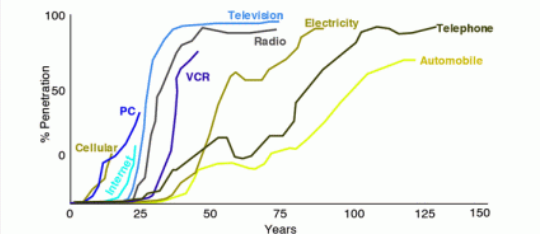
- Internet/Web *evolved* - it wasn't created
- Scalability beats structure
 - search engines over directories
 - Web over hypertext
- "We are 10 seconds from the Big Bang"
 - John Doerr

10/1/2009 4:59 PM 39

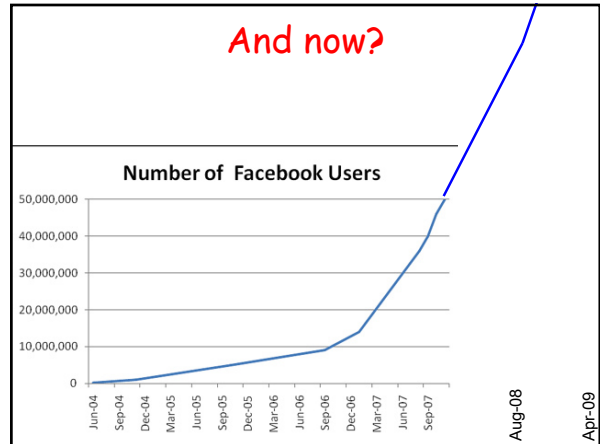
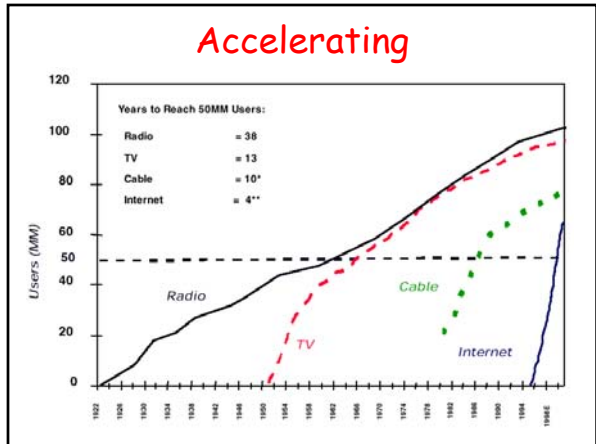
Adoption

Facilitating Innovation the pace of innovation is increasing

- Newer technologies taking hold at double or triple previous rates



10/1/2009 4:59 PM 39

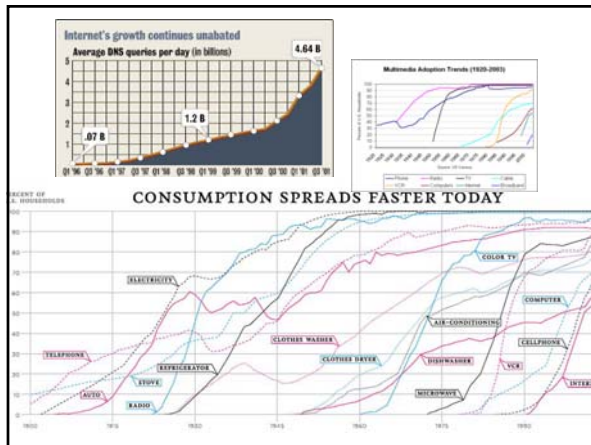
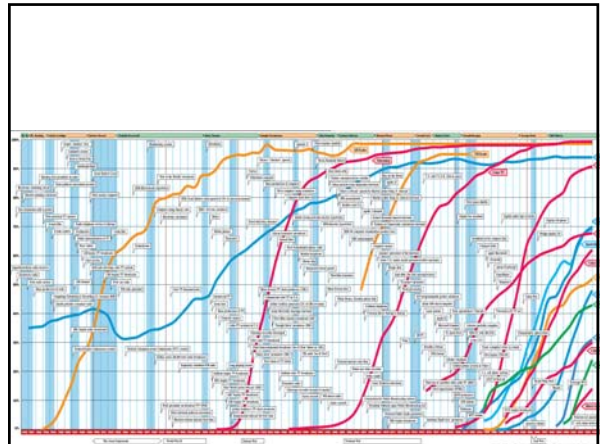


For Next Time

- **Add yourself to mailing list**
 - We'll send out a key email tomorrow
 - Be sure to get it!
- **Think about ps1**
 - Form a group of 2 people
- **Think about project**
 - Form a group of 4 people

10/1/2009 4:59 PM

43



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 Matrix

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%

10/1/2009 4:59 PM

46