

CSE 454 Advanced Internet & Web Services



CSE 454 Advanced Internet & Web Services

- **Prof: Dan Weld**
 - Most lectures, perspective, project org
- **TA: Xiao Ling**
 - Project & code details
- **Expectations:**
 - Project (multiple parts, **on time!**)
 - Reading (papers, web - no formal text)
 - Class participation / development
- **Caveat: Life on the cutting edge**



1/8/2013 4:40 PM

4

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
 - DARPA Machine Reading Project



1/8/2013 4:40 PM

5

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**



Your Background?

- **Year in Program?**
- **Classes?**
 - 444, 446, 451, 461, 473, 490H
- **Concepts?**
 - Race condition?
 - Naïve Bayes classifier?
 - Hybrid hash join algorithm?
 - Precision, recall?
- **Programming Background?**
 - Ruby,
 - .NET,
 - admin own web/db/game server?

1/8/2013 4:40 PM

7

454 Topics

- **Search Engines**
 - Crawling, Indexing, Information Retrieval,
 - Query Processing, Ranking, Pagerank, Interfaces
- **Text Categorization & Clustering**
- **Information Extraction**
 - Machine Learning
 - Natural Language Processing
- **Security, Cryptography, Malware**
- **Human Computation & Social Systems**
- **Internet Advertising & Biz Models**



Today's Outline

- Overview
- Internet: Past & Future
- Class Project & Mechanics

1/8/2013 4:40 PM 10

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.

1/8/2013 4:40 PM 11

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
- 1972 First email programs, FTP spec

1/8/2013 4:40 PM 12

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

1/8/2013 4:40 PM 13

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

1/8/2013 4:40 PM 14

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?

1/8/2013 4:40 PM

15

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

1/8/2013 4:40 PM

16

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: Google, pagerank algorithm
Paypal founded

1/8/2013 4:40 PM

18

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

1/8/2013 4:40 PM

18

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
- 2008: EC2 introduced
- 2009: Facebook 200M users



1/8/2013 4:40 PM


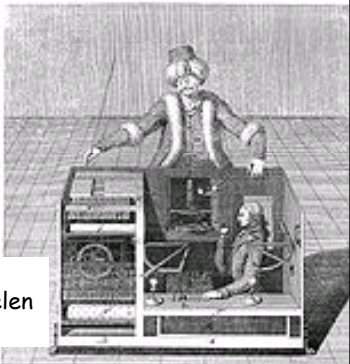
Future of the Net

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

1/8/2013 4:40 PM


20

Mechanical Turk





Built in 1770 by Wolfgang von Kempelen

1/8/2013 4:40 PM



- 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



1/8/2013 4:40 PM 22

Death of the Web

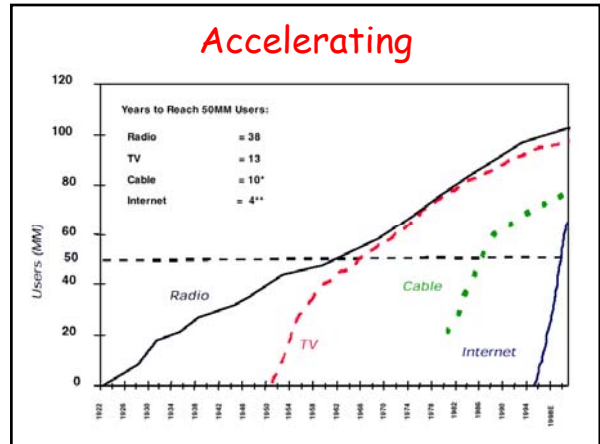
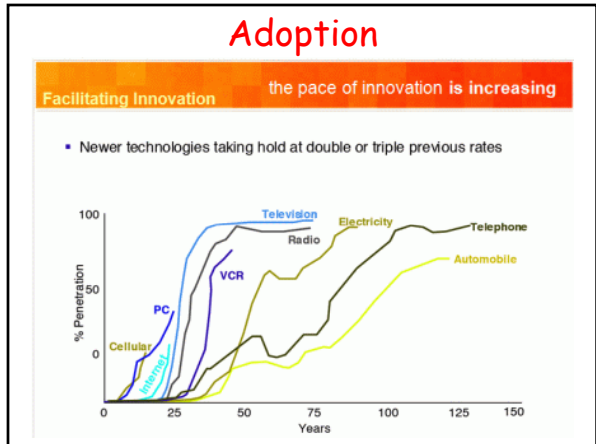
- Pages vs Apps
 - Can't search apps
 - Still use HTTP, but closed protocols

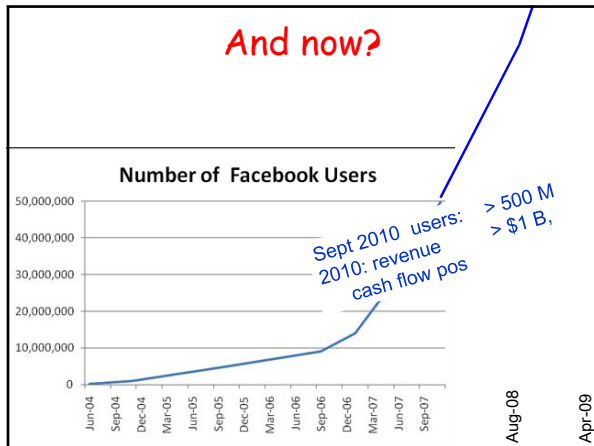
1/8/2013 4:40 PM

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

1/8/2013 4:40 PM 24





- ## Today's Outline
- Overview
 - Internet: Past & Future
 - Class Project & Mechanics
- 1/8/2013 4:40 PM 30

- ## 454 Topics
- Search Engines
 - Crawling, Indexing, Information Retrieval,
 - Query Processing, Ranking, Pagerank, Interfaces
 - Text Categorization & Clustering
 - Information Extraction
 - Machine Learning
 - Natural Language Processing
 - Security, Cryptography, Malware
 - Human Computation & Social Systems
 - Internet Advertising & Biz Models
- 1/8/2013 4:40 PM 32

- ## Why Search?
- Many billions of searches per day...
 - Boost to productivity
 - Intellectual & economic
 - Search is *(still)* 'hot'
 - Amazon, LinkedIn, Yelp, Netflix, Kayak, Maps
 - Fascinating research problem.
 - Yet... you can learn to be a something of a search expert in one quarter!
- 1/8/2013 4:40 PM 32

- ## The Future of Search?
- It's only been 5 min after the big bang...
- 1/8/2013 4:40 PM 33

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 rallied 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 Veighte**, 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 Veighte	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, Libra
 - Product search
- **Question Answering**

1/8/2013 4:40 PM

36

Example

The screenshot shows a Zoominfo search interface. The search criteria are "Person Name" and "Keyword/Company". The results list several entries for Daniel S. Weld, including his role as Venture Partner at Madrona Venture Group LLC, Associate Editor at AI Access Foundation, and various academic and industry affiliations.

1/8/2013 4:40 PM

37

...Continued

The screenshot shows a detailed profile for Dr. Daniel S. Weld. It includes his title as "Venture Partner" at "Madrona Venture Group LLC", a "Contact this person" button, and a list of references and affiliations. The profile also mentions his role as a "Member, Computer Science and Engineering Department" at the University of Washington and his involvement in various research and advisory boards.

1/8/2013 4:40 PM

38

CiteSeer vs. Scholar

The screenshot compares search results for "Daniel Weld" on CiteSeer and Google Scholar. CiteSeer results include titles like "A Softbot-Based Interface to the Internet" and "UCPOP: A sound, complete, partial...". Google Scholar results include titles like "A Scalable Comparison, Abstraction, and..." and "The World-Wide-Web is less agent-friendly...".

Products

The screenshot shows an Amazon product page for "Pre-de Provence Soap - Ounce Cello Wrap". The product is priced at \$7.34 (27% off the list price of \$10.00). It is described as a "250g" soap from France. The page includes the Amazon Prime logo, search bar, and product details.

1/8/2013 4:40 PM

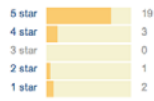
40

...and their Reviews



Customer Reviews

★★★★★ (25)
4.4 out of 5 stars



See all 25 customer reviews

"It leaves your skin feeling very soft and clean."
Amanda C | 9 reviewers made a similar statement

"Scent is mild but very nice and sweet."
Autumn Hays | 6 reviewers made a similar statement

"I can tell that the bar will last a long time compared to other soaps."
ShirlyRand | 6 reviewers made a similar statement

1/8/2013 4:40 PM

41

News...



WASHINGTON — Risking a potentially rancorous battle with Congress at the start of his second term, President Obama on Monday nominated Chuck Hagel, a former Republican senator from Nebraska whom Mr. Obama hailed as "the leader that our troops deserve," to be secretary of defense.

1/8/2013 4:40 PM

42

NIST KBP Challenge

Part 1 - Named Entity Linking

1/8/2013 4:40 PM

43

Part 2 - Slot Filling

Person	Organization
per:alternate names	org:alternate names
per:date of birth	org:political/religious affiliation
per:age	org:top_members/employees
per:country of birth	org:number of employees/members
per:stateorprovince of birth	org:members
per:city of birth	org:member of
per:origin	org:subsidiaries
per:date of death	org:parents
per:country of death	org:founded by
per:stateorprovince of death	org:founded
per:city of death	org:dissolved
per:cause of death	org:country of headquarters
per:countries of residence	org:stateorprovince of headquarters
per:stateorprovinces of residence	org:city of headquarters
per:cities of residence	org:shareholders
per:schools attended	org:website
per:title	

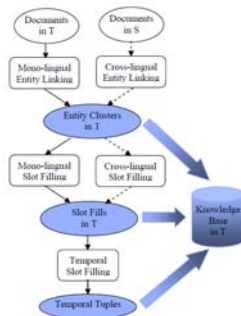
26

16

1/8/2013 4:40 PM

44

2011 Competition Tracks



1/8/2013 4:40 PM

46

Your Mission Project...

- As a class, build an NER/slot-filling system
- Working in small teams
- Combine parts together

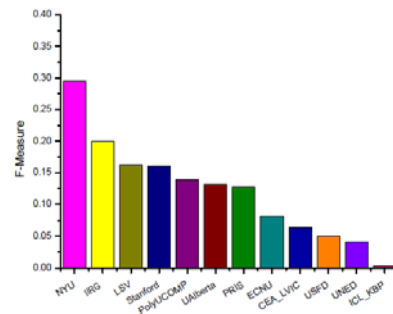
Grading

- 85% Project
 - Part artifact
 - What you did
 - How well it worked
 - Part writeup
 - Clear and concise explanation / justification
 - Experimentation
 - Part presentation
- 15% Class participation

1/8/2013 4:40 PM

47

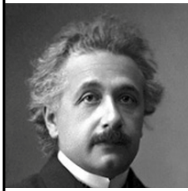
2011 Slot-filling Results



Challenges (Example: Birthdate)

Query: A person name (N)
with some disambiguating context

Attempt: Find sentences "(N) was born in/on (T)"



Albert Einstein was born at Ulm, in Württemberg, Germany, on March 14, 1879. Six weeks later the family moved to Munich, where he later on began his schooling at the Luitpold Gymnasium. Later, they moved to Italy and Albert continued his education at Aarau, Switzerland and in 1896 he entered the Swiss Federal Polytechnic School in Zurich to be trained as a teacher in physics and mathematics. In 1901, the year he gained his diploma, he acquired Swiss citizenship and, as he was unable to find a teaching post, he accepted a position as technical assistant in the Swiss Patent Office. In 1905 he obtained his doctor's degree.

During his stay at the Patent Office, and in his spare time, he

Challenges (Example: Birthdate)

More Real Data:

1. A document mentions a celebrity's 40th birthday is today. Needs to infer her birthdate is 40 years before the publication date.
2. A biography only mentions the person's name in the title. e.g. "Born: Sept 5th, 1951"
3. Other patterns: "She gave birth to David on X"

Preprocessed Data Files

Each line corresponds to a sentence. "John likes eating sausage."

tokens	after tokenization
	John likes eating sausage.

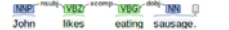
Preprocessed Data Files

Each line corresponds to a sentence. "John likes eating sausage."

tokens	after tokenization
	John likes eating sausage.
pos	Part-of-Speech tags
	John/NNP likes/VBZ eating/VBG sausage/NN ./.

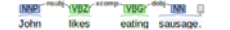
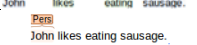
Preprocessed Data Files

Each line corresponds to a sentence. "John likes eating sausage."

tokens	after tokenization	John likes eating sausage.
pos	Part-of-Speech tags	John/NNP likes/VBZ eating/VBG sausage/NN ./.
parse	automatic analysis of grammatical structure	<pre> S (NP (NNP John)) (VP (VBZ likes) (S (VP (VBG eating) (NP (NN sausage)))))) (. .) </pre> *stored in one line
dep	Grammatical dep.	


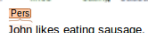
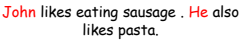
Preprocessed Data Files

Each line corresponds to a sentence. "John likes eating sausage."

tokens	after tokenization	John likes eating sausage.
pos	Part-of-Speech tags	John/NNP likes/VBZ eating/VBG sausage/NN ./.
parse	automatic analysis of grammatical structure	<pre> S (NP (NNP John)) (VP (VBZ likes) (S (VP (VBG eating) (NP (NN sausage)))))) (. .) </pre> *stored in one line
dep	Grammatical dep.	
ner	Named Entities	

Preprocessed Data Files

Each line corresponds to a sentence. "John likes eating sausage."

tokens	after tokenization	John likes eating sausage.
pos	Part-of-Speech tags	John/NNP likes/VBZ eating/VBG sausage/NN ./.
parse	automatic analysis of grammatical structure	<pre> S (NP (NNP John)) (VP (VBZ likes) (S (VP (VBG eating) (NP (NN sausage)))))) (. .) </pre> *stored in one line
dep	Grammatical dep.	
ner	Named Entities	
coref	Coreference	

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!

1/8/2013 4:40 PM

61