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

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

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?

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

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
  - 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.

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

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

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
**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?

**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 started (90% mkt share → 1%)
- **1995:** Ebay
  - MetaCrawler (1st major meta-SE)
    - UW Master’s thesis by Erik Selberg

**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

**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

**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

**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?
Mechanical Turk

Built in 1770 by Wolfgang von Kempelen

Amazon Mechanical Turk

- 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

Death of the Web

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

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

- Newer technologies taking hold at double or triple previous rates

Accelerating

- Years to Reach 500M Users:
  - Radio = 30
  - TV = 15
  - Cable = 10
  - Internet = 8
And now?

Today’s Outline

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

The Future of Search?

- It's only been 5 min after the big bang...

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

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!

What is “Information Extraction”?

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

The Future of Search?

- It's only been 5 min after the big bang...

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

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!

What is “Information Extraction”?

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

The Future of Search?

- It's only been 5 min after the big bang...
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 openly 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...
...and their Reviews

Customer Reviews

“It leaves your skin feeling very soft and clean.”
Amanda Q | 9 reviews made a similar statement

“Sent in mild but very nice and sweet.”
Alina Mays | 6 reviews made a similar statement

“I can tell that the bar will last a long time compared to regular soaps.”
Emily R | 6 reviews made a similar statement

NIST KBP Challenge

• Part 1 – Named Entity Linking

2011 Competition Tracks

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

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)"

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.”

<table>
<thead>
<tr>
<th>tokens</th>
<th>after tokenization</th>
<th>pos</th>
<th>Part-of-Speech tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>likes</td>
<td>NNP</td>
<td>VBG</td>
</tr>
<tr>
<td>eating</td>
<td>sausage</td>
<td>VBZ</td>
<td>VBG</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>./</td>
</tr>
</tbody>
</table>
### Preprocessed Data Files

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

<table>
<thead>
<tr>
<th>tokens after tokenization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pos</strong> Part-of-Speech tags</td>
</tr>
</tbody>
</table>
| John/NNP  
likes/VBZ  
eating/VBG  
sausage/NN ./, |  
| **parse** automatic analysis of grammatical structure |  
| *stored in one line |  
| **dep** Grammatical dep. |  
| John likes eating sausage. |  
| **ner** Named Entities |  
| John likes eating sausage. |  
| **coref** Coreference |  
| John likes eating sausage, He also likes pasta. |