CSE 454
Advanced Internet & Web Services

Oren Etzioni
Paul Pham

(slides from Dan Weld, Tom Alberg, and others)
Your Background?

• Year in Program?

• Classes?
  - 444, 446, 473, 461

• Programming Background?
  - Web programming?
  - Android programming?
  - “cool” Internship?
My Background

• Research on AI, ML, and NLP.
• Was the CTO of a “top 10” Internet company around 2000.
• Start companies
  - 1996: Netbot: first comparison-shopping.com
  - 2003: Farecast “Know when to buy” airfares
  - 2010: Decide.com “No regrets” in CE

• Haven’t written code for a long time.
Outline

1. Search (motivation/background for 454)
2. Quick overview of RevMiner
3. Project
4. Administrivia
1. Search (past, present, future)
Why Search?

• A billion or so searches per day...
• Boost to productivity
  - Intellectual & economic
• Search is important
  - Google, Bing, Amazon, Yelp, Match.com
  - Search for products, music, people, ...
• Fascinating research problem
• You can learn to be a something of a search expert in one quarter!
• 454 will help you be a great job candidate!
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 (product search)

• **1995**: MetaCrawler (1st major meta-SE)
  - UW Master’s thesis by Erik Selberg
Discovery of the Biz Model

1997: goto.com
  “sponsored links” pay-per-click

AskJeeves
  manually-powered question answering

Netbot
  comparison-shopping search

1998: Open directory launched
  Google, pagerank algorithm

• 1999: Search Engines → portals (Yahoo, Excite)
• 2001: Ascendence of Google
Why Did Google Win?

• Fast
• Clean, uncluttered UI
• Better ranking based on pagerank, anchor text, other tricks
• Key partnerships (AOL, Mozilla, others)
• Better ad targeting
Approaching the Present

• 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
• 2011: Bing re-emerges...

Facebook (800M UU) threat?
Future of Search

- Link-Spamming (Arms race to bias SE ranking)
- Local Search
- Image & Video search
- What else?
- Mobile Devices (cellphone, etc)
Is Search Solved?

- So no more fun to be had...
- Are you happy with search?
- How about search on your phone?
Explosive growth of Mobile Devices
Mobile, Mobile, Mobile, Mobile!
Cell Phone Innovation Continues...
What Technologies Enable “Mobile”?

- Flash Storage
- Fast Bandwidth
- Touch
- Powerful
- Voice Technology
- Cloud Resources
- Fast Processors
Wireless Data Speeds are Increasing

Source: Dell ‘Oro, and Goldman Sachs Research estimates.
Mobile is becoming Primary!

% of Traffic From Mobile Devices, Pandora, Twitter & Facebook, 2008 – 2011

Source: Mary Meeker Internet Trends Presentation, 10/18/11
Focus of CSE 454 in 2012

- Search from a mobile device
- Googlers kindly donate Android phones
- Project Focus: build the best mobile restaurant search!
- Building blocks:
  - Restaurant meta-data
  - 400,000 yelp reviews
  - RevMiner
Provocative Comments...

“We could soon view today’s keyword searching with the same amusement and nostalgia reserved for...electric typewriters and vinyl records.”

(Etzioni, 8/4/11)
Search ≠ Google circa 2008

- Search-from-phone is the future
  - Local: restaurants, movies, flights
  - Social: Facebook, Twitter
  - Shopping: Milo, Decide.com

- Impressive progress in semantic search & Natural Language Processing (NLP)
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…
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…

<table>
<thead>
<tr>
<th>NAME</th>
<th>TITLE</th>
<th>ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>founder</td>
<td>Free Soft.</td>
</tr>
</tbody>
</table>

Slides from Cohen & McCallum
Paradigm Shift: from retrieval to extraction

KnowItAll

How is the iPad 2?

World Wide Web

Found 28,900 reviews; 87% positive.

Key features include...

Google

Information Food Chain

26
2. RevMiner
RevMiner (Huang, Etzioni, Zettlemoyer)

- Extracts key attributes + opinions
- Applied to Yelp reviews (Seattle)
- Based on Opine (Popescu & Etzioni '05)

Extractive UI versus search UI
(Yatani et. al, CHI '11)
RevMiner is a University of Washington KnowItAll project by Jeff Huang, Oren Etzioni, Luke Zettlemoyer

RevMiner is an unsupervised extractor for user reviews about places in Seattle, like restaurants and stores

Try these examples:
good dim sum
aguava verde cafe
cheap indian food
great view clean rooms comfortable bed
great location
elliott bay book
mexican
free parking friendly staff fresh fish
<table>
<thead>
<tr>
<th>Category</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>sushi</td>
<td>freshest, incredible, superb, amazing, fantastic, awesome, perfect, favorite, reasonable, excellent, best, delicious, affordable, fancy, worth, warm, inexpensive, fresh, traditional, not solid, tasty, huge, top, prepared, quick, real, korean, different, full, japanese, not bad, decent, expensive, not great, not good, raw, average, bad, ok, alright, ok, not fresh, mediocre, poor</td>
</tr>
<tr>
<td>place</td>
<td>freshest, amazing, superb, awesome, glad, favorite, perfect, intimate, delicious, gorgeous, excellent, best, yummy, classy, great, enormous, simple, worth, chic, everyday, authentic, fresh, impressive, fast, top, fun, tasty, modern, not loud, clean, entertaining, solid, fancy, sour, pricier, cute, full, not small, crazy, different, packed, loyal, latest, central, hot, good, big, early, bigger, small, nice, japanese, study, price, tiny, open, long, popular, specific, close, tired, late, busy, cheap, not top, last, trendy, typical, not damn, giant, higher, not free, bad, decent, complete, excited, empty, overpriced, poor, rude, not nasty, not stuffy</td>
</tr>
<tr>
<td>rolls</td>
<td>wonderful, amazing, awesome, fantastic, perfect, delicious, best, excellent, creative, yummy, great, beautiful, super, unique, not unique, huge, light, tasty, large, extra, japanese, good, different, interesting, hot, full, big, fried, small, nice, not better, not standard, standard, decent, fine, not tasty, raw, not good, average, okay, bad, not mediocre</td>
</tr>
<tr>
<td>food</td>
<td>fabulous, outstanding, amazing, wonderful, fantastic, awesome, pure, favorite, professional, excellent, affordable, delicious, best, yu, friendly, fresh, not sweet, inexpensive, worth, generous, classic, authentic, baked, consistent, not delicious, expensive, damn, fast, top, tasty, spicy, regular, spicily, full, japanese, smaller, cheap, late, enough, whole, pricey, usual, memorable, typical, not hard, not familiar, standard, bad, hungry, half, par, fine, slow, disappointed, cold, average, bad, okay, ok, alright, not impressed, mediocre, passable, overpriced, subpar, worst</td>
</tr>
<tr>
<td>service</td>
<td>incredible, impeccable, amazing, wonderful, superb, fabulous, outstanding, awesome, fantastic, reasonable, helpful, excellent, excellent, friendly, efficient, stellar, warm, speedy, attentive, top, fast, prompt, consistent, quick, large, solid, polite, full, above, good, best, usual, not old, standard, better, not consistent, spotty, decent, not great, fine, par, not hot, not attentive, slow, average, pretentious, bad, sucked, not mediocre, inattentive, poor, awful, terrible, horrible, worse</td>
</tr>
<tr>
<td>happy hour</td>
<td>incredible, fabulous, amazing, awesome, fantastic, marvelous, wonderful, unbeatable, excellent, delicious, best, reasonable, impressed, great, happy, different, low, good, big, regular, nice, late, long, busy, cheap, last, not bad, better, decent, not good, bad, bad, ridiculious, not magnificent</td>
</tr>
<tr>
<td>atmosphere</td>
<td>incredible, fabulous, amazing, wonderful, fantastic, perfect, wonderful, delicious, not excellent, relaxed, best, romantic, lovely, unique, classy, comfortable, lively, modern, fun, contemporary, attentive, cute, large, cool, coddling, exciting, pleasant, fancy, swanky, good, different, not good, nic, not small, bad, not terrible, sucks</td>
</tr>
<tr>
<td>prices</td>
<td>incredible, awesome, reasonable, not cranky, great, worth, fair, huge, moderate, steeper, sweet, large, small, regular, full, spendy, not cheap, steep, high, higher, hard, typical, tricky, sticky, due, better, not bad, lower, standard, decent, expensive, not sticky, average, acceptable, dry, overcooked, wet</td>
</tr>
<tr>
<td>fish</td>
<td>freshest, amazing, awesome, delicious, best, excellent, reasonable, flavorful, super, great, thing, fresh, generous, prepared, huge, ideal, top, tasty, good, high, better, not fishy, not nice, fine, fishy, average, bad, not fresh, mushy</td>
</tr>
<tr>
<td>server</td>
<td>knowledgeable, wonderful, amazing, awesome, best, adorable, helpful, friendly, patient, super, gracious, great, efficient, pleasant, quick, different, nice, good, busy, long, whole, not nice, not bad, not knowledgeable, fine, slow, not friendly, not helpful, terrible, not bad, not solid, not helpful, not fishy, not nice, not okay, not average, not bad, not good, not perfect, not fantastic, not delicious</td>
</tr>
<tr>
<td>menu</td>
<td>incredible, amazing, fantastic, delicious, creative, unique, super, great, inventive, diverse, enormous, wide, impressive, fresh, tasty, huge, large, robust, able, nice, full, big, regular, good, crazy, interesting, small, long, normal, not traditional, typical, special, standard, decent, fine, not poor, not terrible, sucks</td>
</tr>
<tr>
<td>restaurant</td>
<td>incredible, amazing, favorite, best, classy, romantic, great, nice, airy, beautiful, worth, authentic, top, spacious, clean, tasty, huge, large, larger, pleased, nice, good, short, small, dark, whole, open, hard, not free, better, front, not packed, loud, fine, average, empty, not busy</td>
</tr>
</tbody>
</table>
Future of RevMiner

• “Squeeze” onto an Android phone.

• Basis for our course to add:
  - Meta-data
  - Pictures
  - Social features
  - “the sky’s the limit”
3. Project Progression

1. Brain storm & find partner (by Thursday)
2. Propose initial idea (1 page, next week)

Project:

1. HTML5 Implementation
2. Android app.
3. Finale/measurements
4. Project presentation (class 5-10 minutes each)
Project Ideas

• **Different UI for mobile**
  - Voice interface
  - Map interface
  - Different ways to navigate

• **Add information**
  - Health code violations (web site)
  - Food pictures (from where?)
  - Menu info. (openmenu.org)
  - What else?
Project Ideas cont.

- **Add capability**
  - Search for dish + review

- **Add social component**
  - Which of my friends ate there? What did they say?
  - Rank reviews by user

- YOUR GREAT IDEA <here>

- (it’s ok to have multiple groups working on same)
4. Administrivia

- **Prof: Oren Etzioni**
  - lectures, concepts, perspective.
- **TA: Paul Pham**
  - Project details
- **Expectations:**
  - Project (multiple parts, **on time**!)
  - Reading (papers, web - no formal text)
  - Class participation / development
- **Caveat: Life on the cutting edge**
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!
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.


• 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
454 Topics

• **Information Retrieval**
• **Search Engines**
  - Crawling, Indexing, Query Processing, Ranking
  - Pagerank, Interfaces
• **Text Categorization & Clustering**
• **Information Extraction**
  - Machine Learning
• **Social Networks**
• **Special Topics**
Project Groups

• **Group of 2 is ideal**
  - Learn to work together
  - Not too much overhead

• **Exceptions have to be approved, and have a compelling reason**
Grading

- **85% Project (Staged in Parts)**
  - Part artifact
  - Part writeup
    - Clear and concise explanation / justification
    - Experimentation
  - Part presentation
  - Each person will have a distinct part, and will be graded separately

- **15% Class participation**
For Next Time

• Add yourself to mailing list

• Brain storm Project possibilities
  - Try Yelp
  - Play with RevMiner
  - What do *you* want?

• Form a group of 2 people