COEN 169

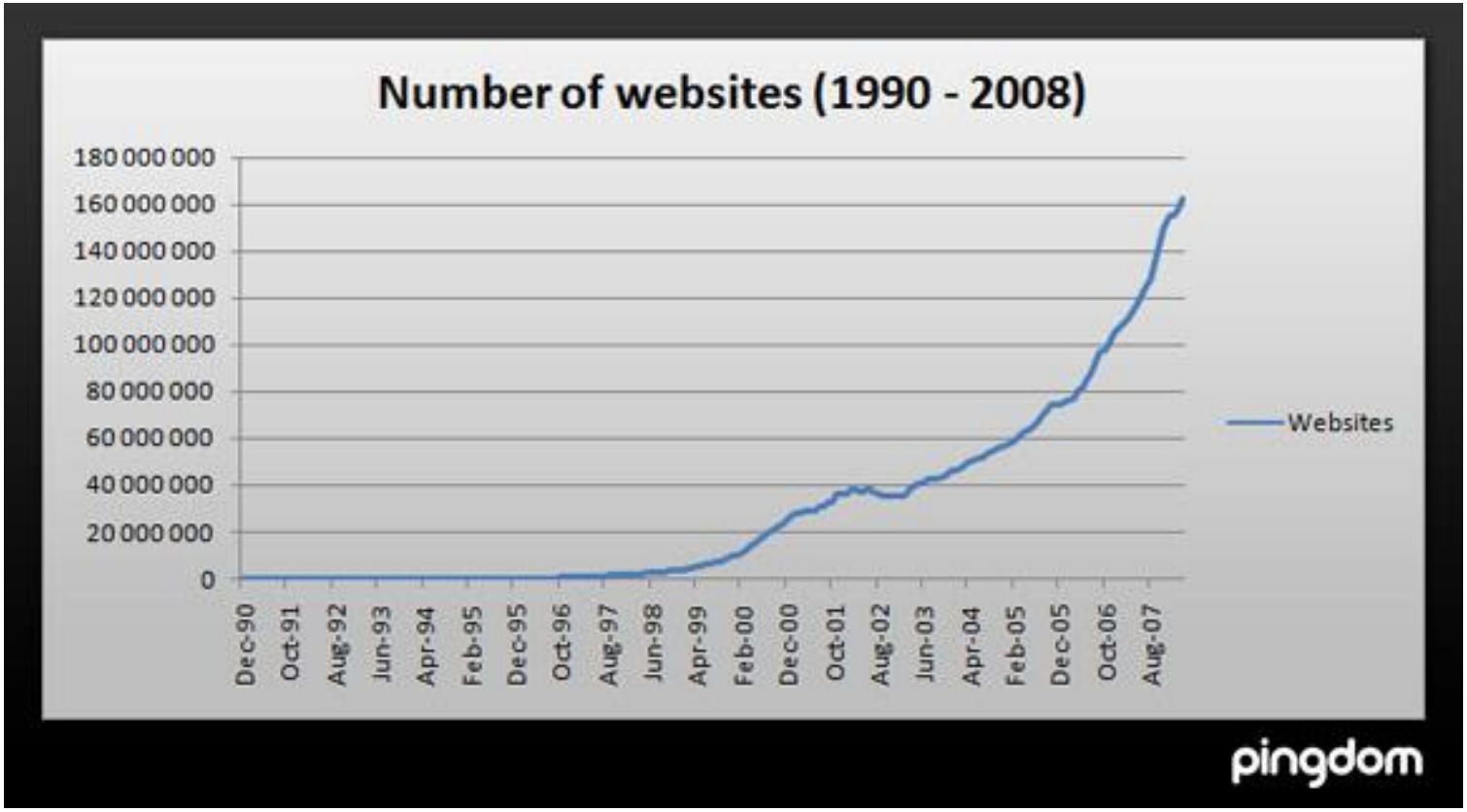
Web Information Management

Yi Fang, Ph.D.
http://www.cse.scu.edu/~yfang

Department of Computer Engineering

Santa Clara University

Web growth


Number of websites (1990 - 2008)
180 000 000
160 000 000
140 000 000
120 000 000
100 000 000
80 000 000
60 000 000
40 000 000
20 000 000
0
Dec-90
Oct-91
Aug-92
Jun-93
Apr-94
Feb-95
Dec-95
Oct-96
Aug-97
Jun-98
Apr-99
Feb-00
Dec-00
Oct-01
Aug-02
Jun-03
Apr-04
Feb-05
Dec-05
Oct-06
Aug-07
Websites
pingdom

Web Information Management

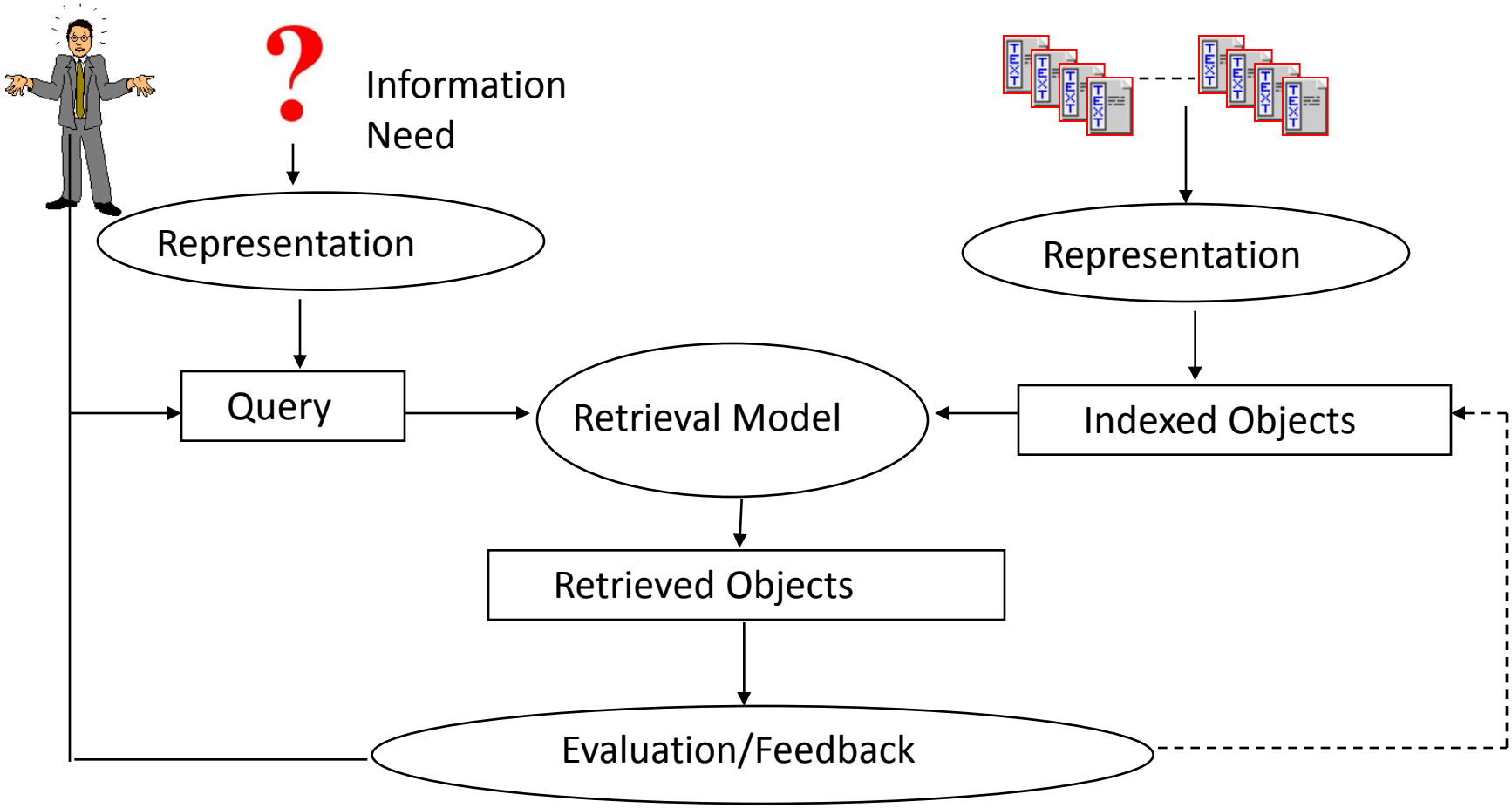
Theory, design, and implementation of information systems that process, organize, analyze large-scale information on the Web

- Search engines
- Recommendation systems

History of Search Engines

Year	Engine	Current status
1993	W3Catalog	Inactive
	Aliweb	Inactive
1994	WebCrawler	Active, Aggregator
	Go.com	Active, Yahoo Search
	Lycos	Active
1995	AltaVista	Inactive (URL redirected to Yahoo!)
	Daum	Active
	Magellan	Inactive
	Excite	Active
	SAPO	Active
	Yahoo!	Active, Launched as a directory
1996	Dogpile	Active, Aggregator
	Inktomi	Acquired by Yahoo!
	HotBot	Active (lycos.com)
	Ask Jeeves	Active (rebranded ask.com)
1997	Northern Light	Inactive
	Yandex	Active
1998	Google	Active
	MSN Search	Active as Bing
1999	AlltheWeb	Inactive (URL redirected to Yahoo!)
	GenieKnows	Active, rebranded Yellowee.com
	Naver	Active
	Teoma	Active
	Vivisimo	Inactive
2000	Baidu	Active
	Exalead	Acquired by Dassault Systèmes

Search Process



Some core concepts

Query Representation:

- Bridge lexical gap: system and systems; create and creating
- Bridge semantic gap: car and automobile

Document Representation:

- Internal representation of document contents: a list of documents that contain specific word
- Representation of document structure: different fields (e.g., title, body)

Retrieval Model:

- Algorithms that best match meaning of user query and available documents. (e.g., vector space model and statistical language modeling)

Applications

Web Information Management: a gold mine of applications

- **Web Search**
- **Recommendation Systems**
- Information Organization: text categorization; document clustering
- Information Extraction: deep analysis of the surface text data
- Entity retrieval and Question-Answering: find the answer directly
- Social network analysis
- Multimedia Information Retrieval: image, video, music, etc.
- Cross-language retrieval
- Information Visualization: Let user understand the results in the best way
- …….

Why WIR or Information Retrieval?

- Information Retrieval (IR) mainly studies unstructured data:

  Text in Web pages or emails; image; audio; video; protein sequences..

Merrill Lynch estimates that more than 85 percent of all business information exists as unstructured data - commonly appearing in e-mails, memos, notes from call centers and support operations, news, user groups, chats, reports, … and Web pages.

IR vs. Database

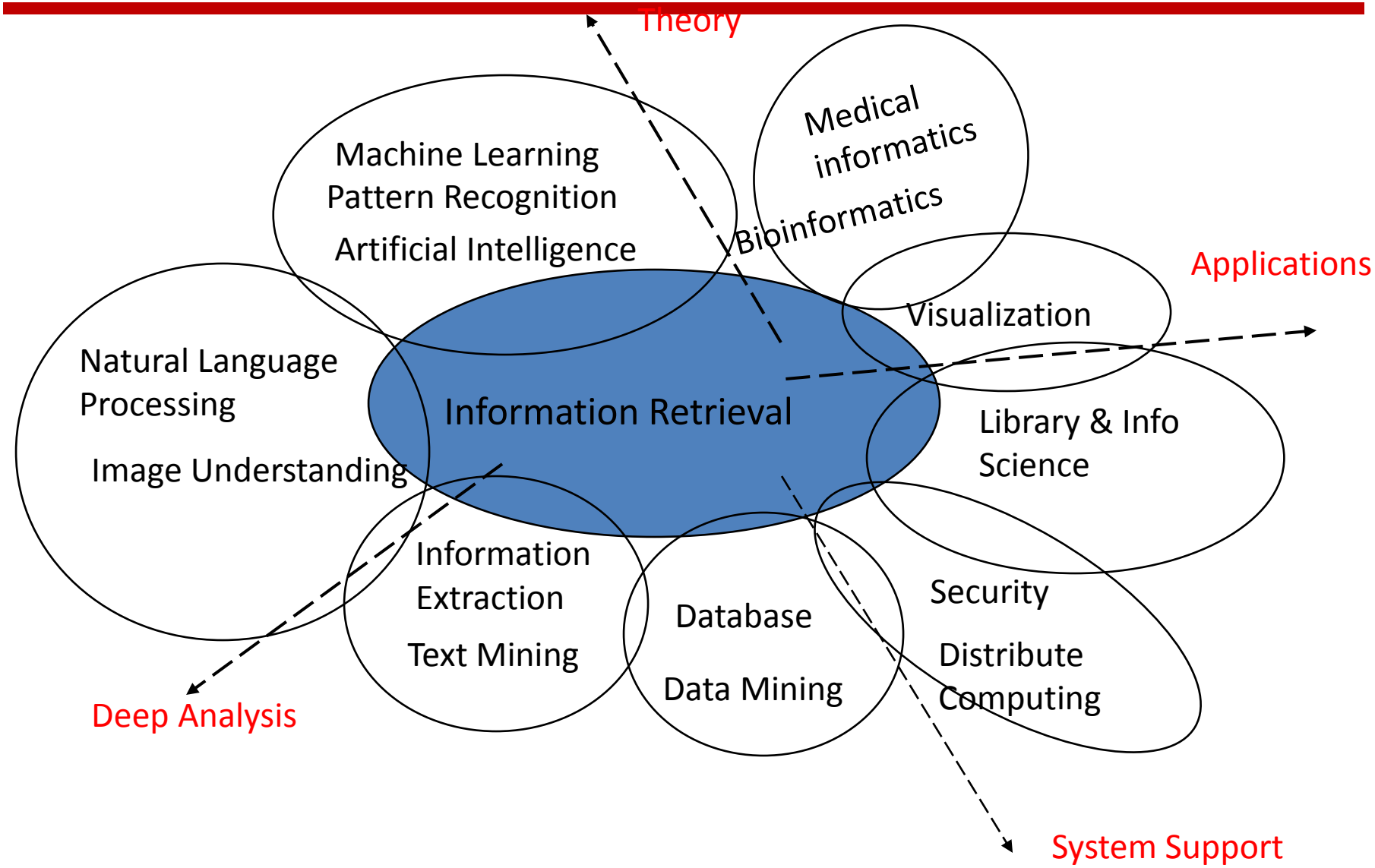
Relational Database Management Systems (RDBMS):

- Semantics of each object are well defined
- Complex query languages (e.g., SQL)
- Exact retrieval for what you ask
- Emphasis on efficiency

Information Retrieval (IR):

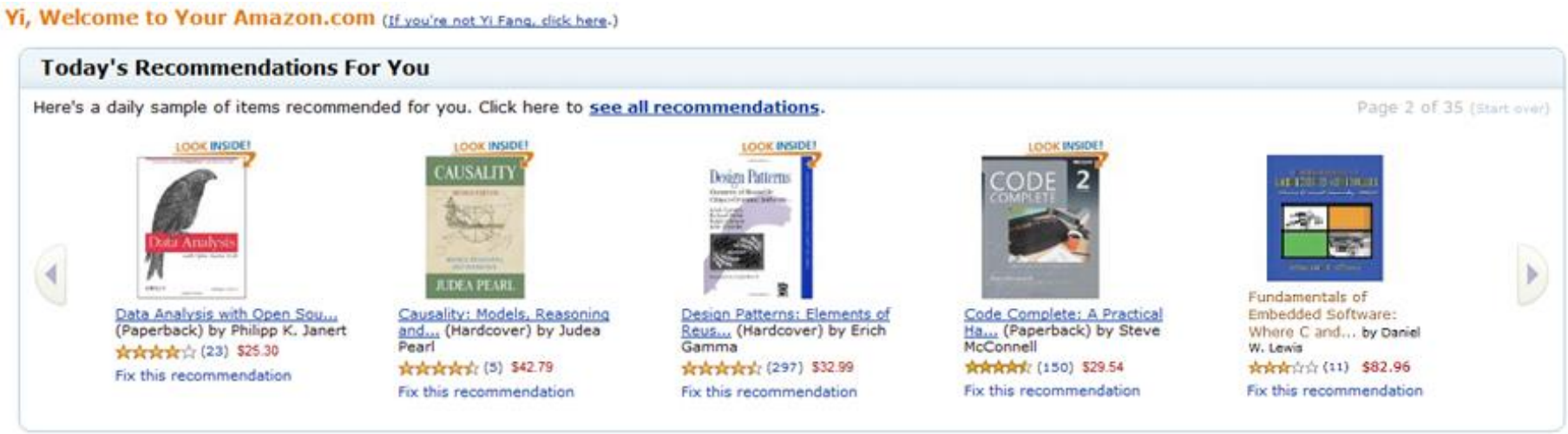
- Semantics of object are subjective, not well defined
- Usually simple query languages (e.g., natural language query)
- You should get what you want, even the query is bad
- Effectiveness is the primary issue, although efficiency is important

IR and other disciplines



Book Recommendation systems

- **Amazon.com** recommends books based on your purchase history (and others')



Movie Recommendation systems

- **Netflix** predicts other "Movies You'll Love"

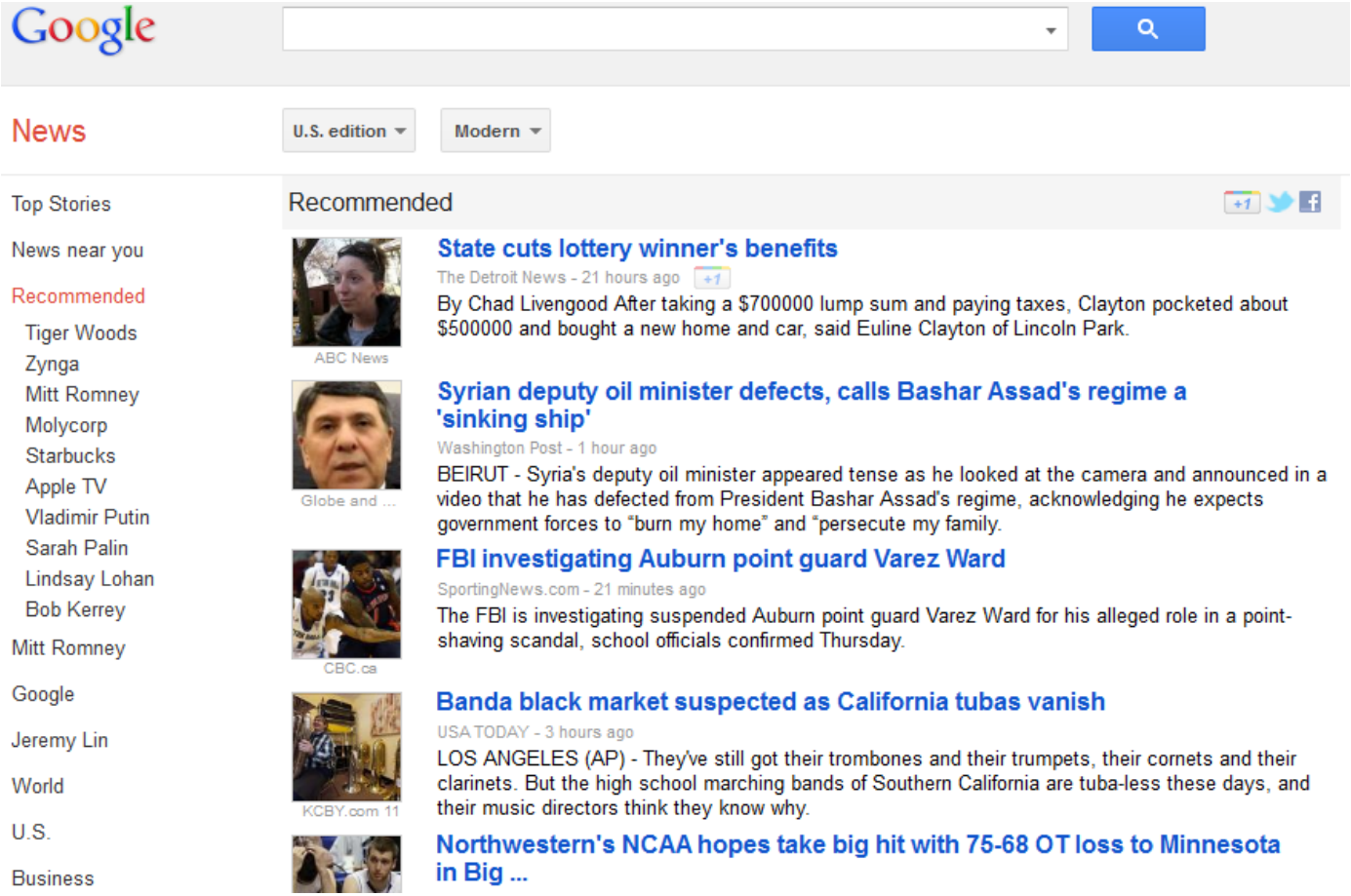
➢ Recommendations drives more than 60% Netflix's DVD rentals [Thompson, 2011]

Recommendation algorithms

- **Netflix Prize:**
  Beat Netflix's own recommender system with 10% margin,
  Win $1 million

- **Testbed:**
  480,000 users
  18,000 movies

News Recommendation



- **Google News** recommends news articles based on clicks and browse history

Personalized Job Recommendation

User Click Prediction in Personalized Job Recommendation

Miao Jiang, Yi Fang
Department of Computer Engineering
Santa Clara University
Santa Clara, California, USA
yfang@scu.edu

Huangming Xie, Jike Chong, Meng Meng
Simply Hired, Inc.
Sunnyvale, California, USA
{jike, huangming}@simplyhired.com

ABSTRACT

Major job search engines aggregate tens of millions of job postings online to enable job seekers to find valuable employment opportunities. Predicting the probability that a given user clicks on jobs is crucial to job search engines as the prediction can be used to provide personalized job recommendations for job seekers. This paper presents a real-world job recommender system in which job seekers subscribe to email alert to receive new job postings that match their specific interests. The architecture of the system is introduced with the focus on the recommendation and ranking component. Based on observations of click behaviors of a large number of users in a major job search engine, we develop a set of features that reflect the click behavior of individual job seekers. Furthermore, we observe that patterns of missing features may indicate various types of job seekers. We

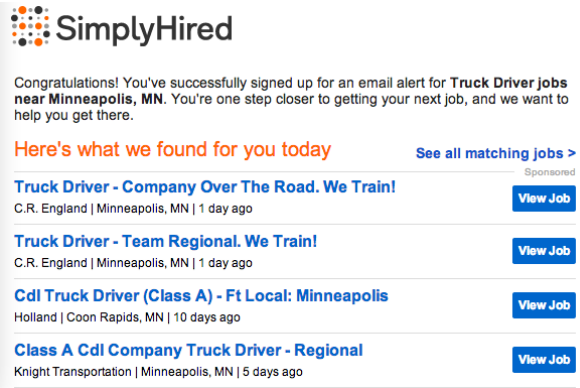


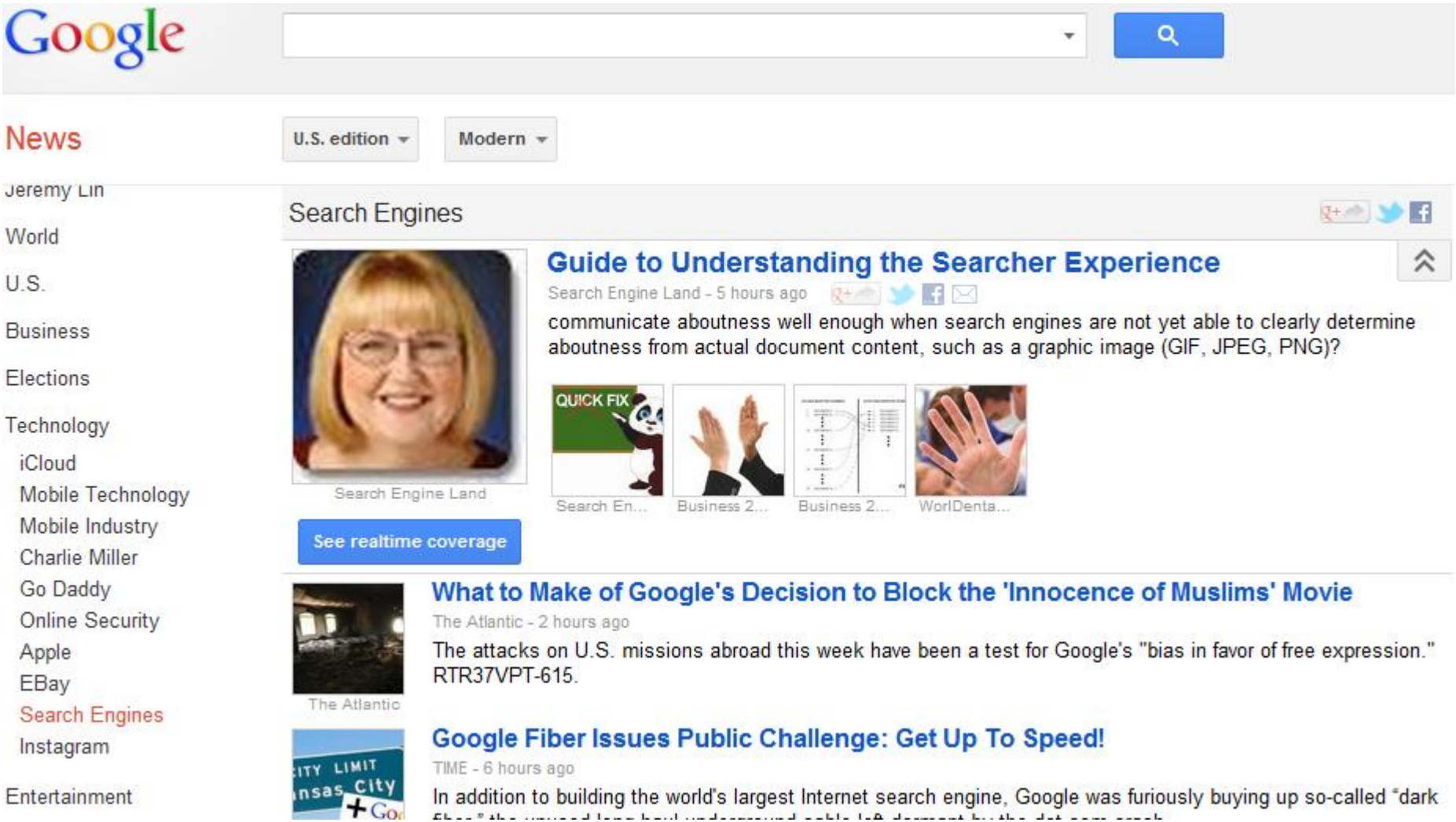
Figure 1: An example of *Simply Hired*'s email alert service for job recommendation.

Point-of-Interest Recommendation

- Foursquare check-in

- Google Place API

IR Applications: Text Categorization



IR Applications: Document Clustering

Clusty

Web+ | News | Images | Shopping | Wikipedia | Blogs | Jobs | Customize!

Java

Cluster by: Topics

All Results (265)

- Software (40)
- Java Technology (25)
- Java.net (28)
- Download (22)
- Book (17)
- Java Applets (18)
- Enterprise (13)
- JavaScript (10)
- Indonesia, Island (8)
- Object, Open Source (7)

more | all clusters

Top 265 results of at least 94,536,601 retrieved for the query Java (Details)

News

- Sun offers Java beta release (Yahoo! News) Jun 21, 2006

java Errors Fixed Free

java errors all Fixed instantly! Free download available Now dllfix.net

Free JSP Editor - Eclipse

BEA Workshop Studio for JSP, Struts JSF IDE, EJB3 - Download now! www.bea.com

1. Java Technology

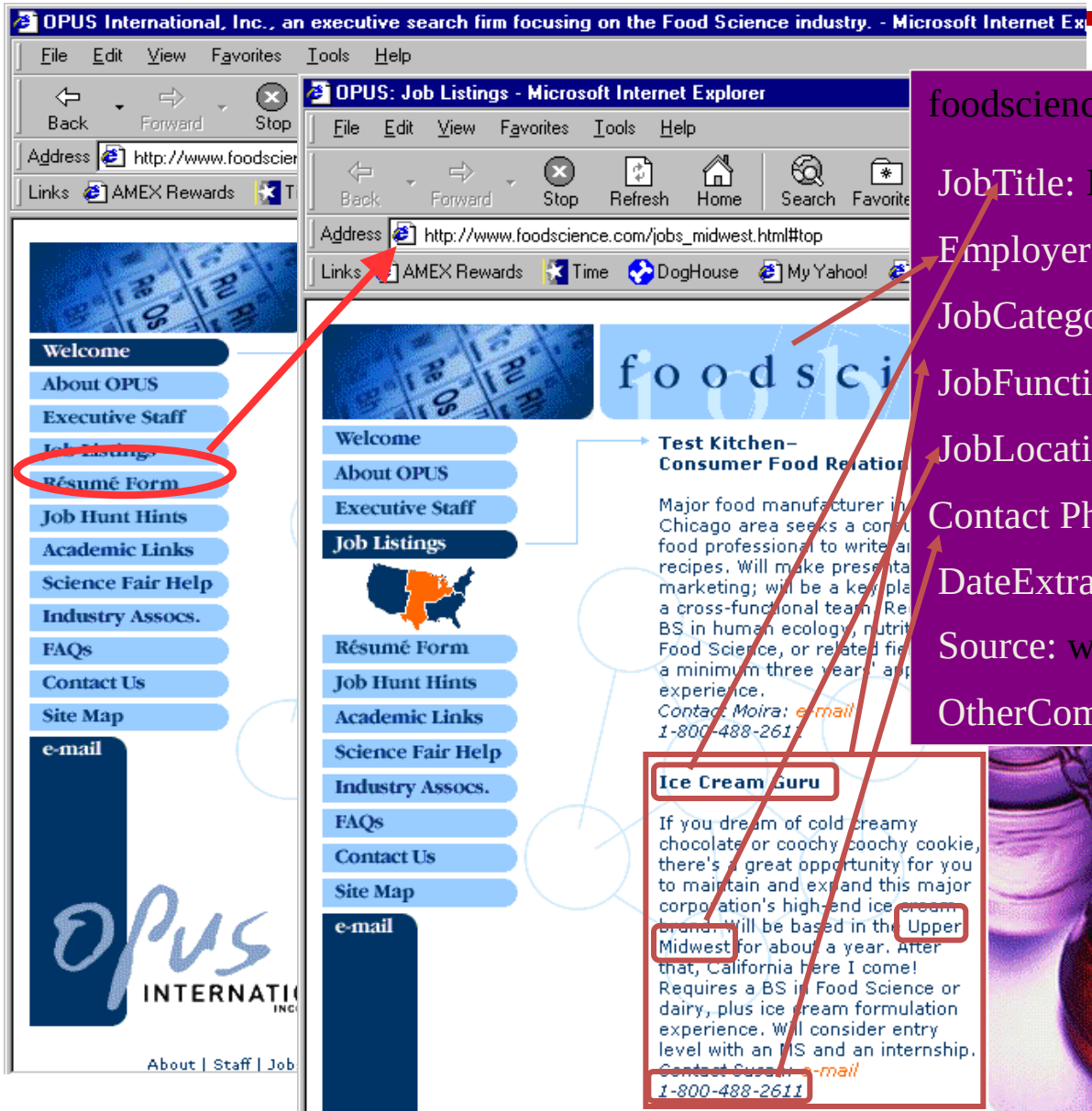
Sun's home for Java. Offers Windows, Solaris, and Linux Java Development Kits (JDKs product information.

java.sun.com - [cache] - Ask, Gigablast, MSN, Wisenut

2. Java programming language

Java is an object-oriented programming language developed initiall colleagues at Sun Microsystems. Initially called Oak (named after office), it was intended to replace C++, although the feature set bet Java should not be confused with JavaScript, which shares only the

IR Applications: Information Extraction



foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.htm

OtherCompanyJobs: foodscience.com-Job1

Entity Search

Google

Universities that are members of the West Coast Conference

Search

About 4,680,000 results (0.18 seconds)

Everything
Images
Maps
Videos
News
Shopping
More

West Lafayette, IN
Change location

Show search tools

The **West Coast Conference** Official Athletic Site
www.wccsports.com/
The The **West Coast Conference** Official Athletic Site, partner of CBS **College** Sports Networks, Inc. The most comprehensive coverage of The West Coast ...

BYU Becomes Ninth **Member** of **West Coast Conference** - West ...
www.wccsports.com/genrel/070111aab.html
Jul 1, 2011 – For the first time in over thirty years, the **West Coast Conference** has a new **member**. Brigham Young **University** formally joins the WCC on ...

West Coast Conference Announces 2012 WCC Hall of Honor Class ...
www.wccsports.com/genrel/021412aac.html
Feb 14, 2012 – The induction ceremony will be **part** of the **Conference's** celebration of its rich history in athletics ... Elaine Michaelis, Brigham Young **University** ...

West Coast Conference adds CSU Bakersfield as affiliate **member** i...
www.wccsports.com/sports/w-golf/spec-rel/120111aaa.html
Dec 1, 2011 – The **West Coast Conference** announced today the addition of California State **University** Bakersfield as an affiliate **member** in the sport of ...

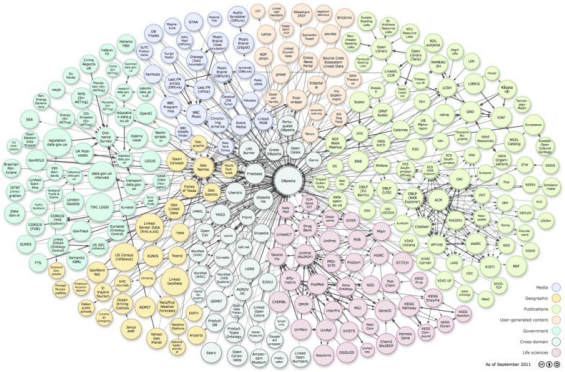
West Coast Conference - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**West_Coast_Conference**
Jump to Former **members**: **University** of the Pacific (1952–1971) (now a **member** of

Why Entity Search?

- User demand

  More than 50% of Web search queries target on entities [Pound et al., WWW10]

- Web science and technology
  - Sophisticated Web mining techniques
  - Data in Semantic Web naturally centered around entities



Expert Search

INDURE: Indiana database of university research

www.indure.org

Question Answering

Google

What companies build parts used in production of Ford vehicles?

Search

About 334,000,000 results (0.16 seconds)

Everything
Images
Maps
Videos
News
Shopping
More

West Lafayette, IN
Change location

Show search tools

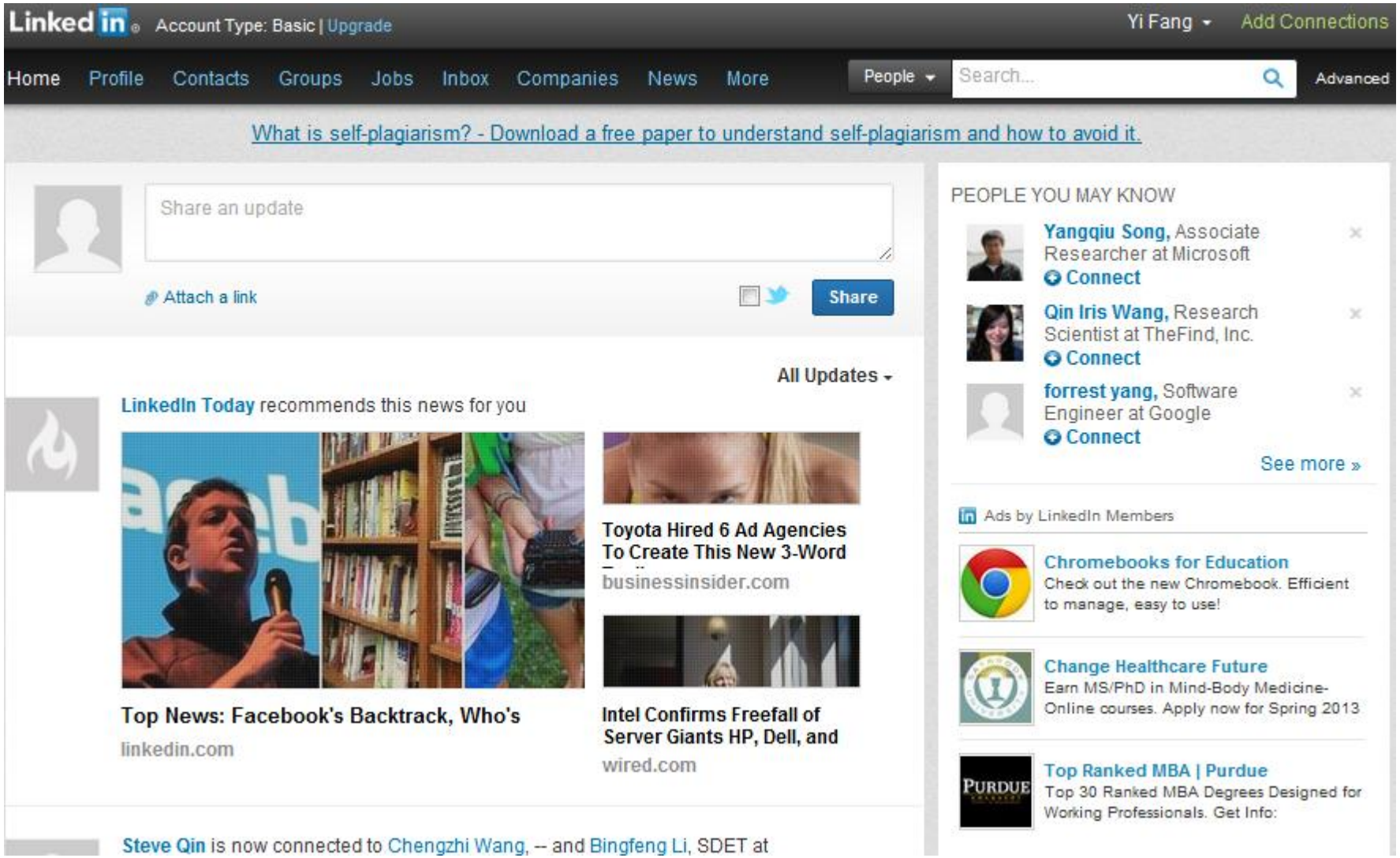
Assembly line - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Assembly_line
Ford was the first **company** to **build** large factories around the assembly line concept. ... (3) **Use** sliding assembling lines by which the **parts** to be assembled are delivered ... In traditional **production**, only one **car** would be assembled at a time.
↳ Concept - History - Sociological problems - Improved working conditions

Ford Motor **Company** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Ford**_Motor_**Company**
During its early years, the **company produced** just a few **cars** a day at its factory ... line concept; and **Ford** soon brought much of the **part production** in-house in a ... In January 2008, **Ford** launched a website listing the ten **Built Ford** Tough rules .... From the 1940s to late 1970s **Ford's Ford** F-Series were **used** as the base for ...

Land Rover - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Land_Rover
In 1967 the Rover **Company** became **part** of Leyland Motor Corporation and in ... **built** under license and the engine was also **used** in **Ford** pick-up **trucks built** locally. **Production** of the TDi engine ended in the United Kingdom in 2006, meaning ...

IBM Watson in *Jeopardy!*

Social Network Analysis



IR Applications: Multimedia Retrieval

Query

Color Histogram

Wavelet…

Feature Extraction

Retrieval Model

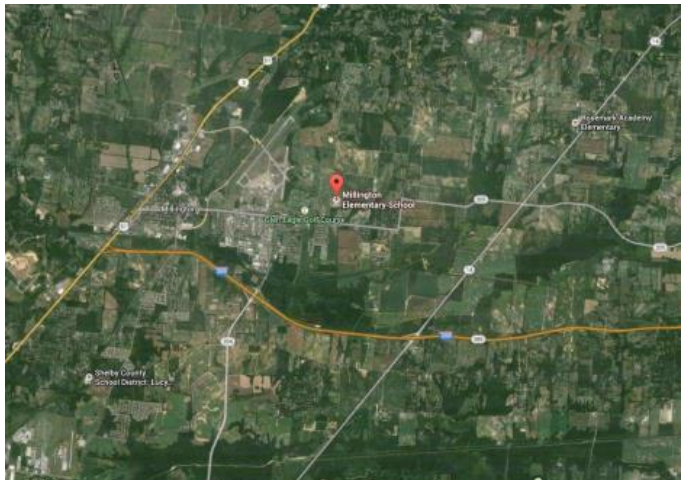
Pictures

Feature Extraction

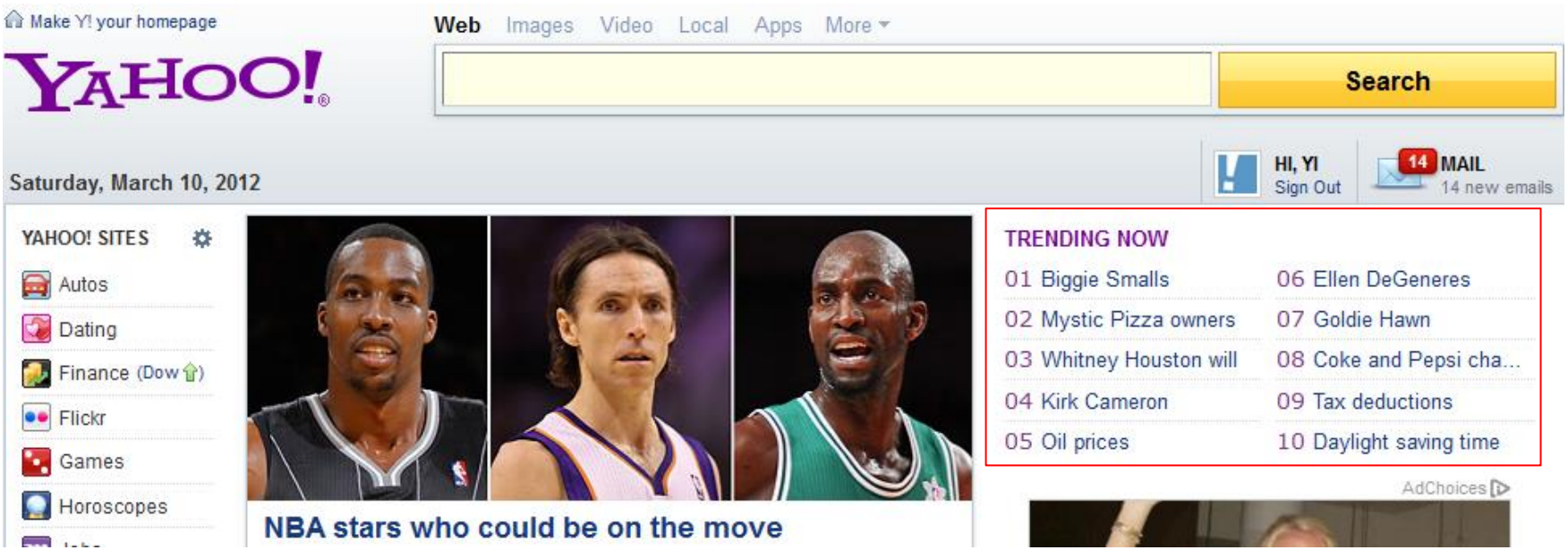
Detecting Road Hazards Using Twitter

Micheal Anderson @manderson1987 · 17m

I brake for deers literally! Saw about 13 deers leaping across the road 3 hr drive in dark dangerous!
twitpic.com/aawx1



Trending topic discovery



- Analyze > 3,200 queries per second
- Cloud computing

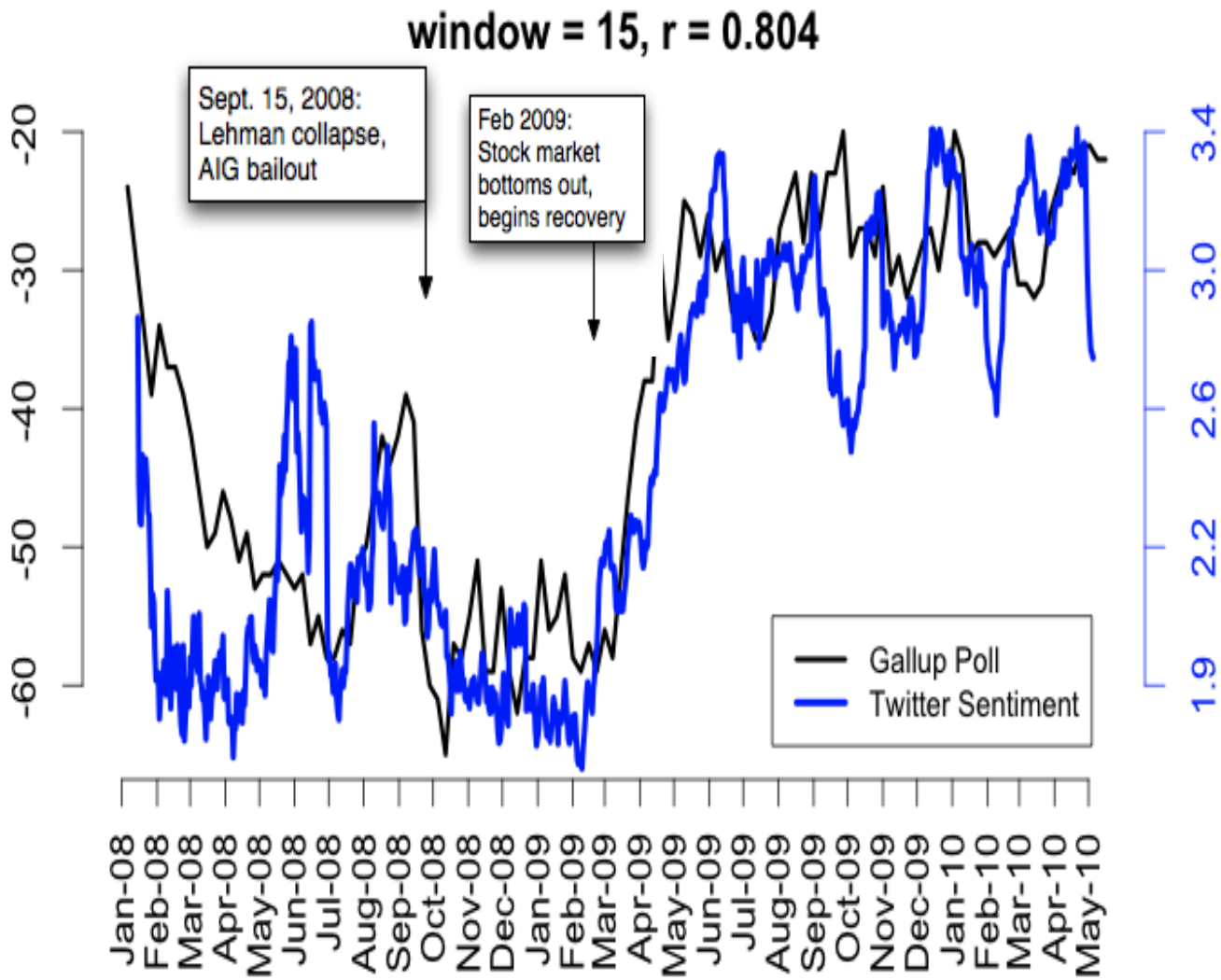
hadoop

Sentiment Analysis

Positive or negative movie review?

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

Twitter sentiment versus Gallup Poll of Consumer Confidence

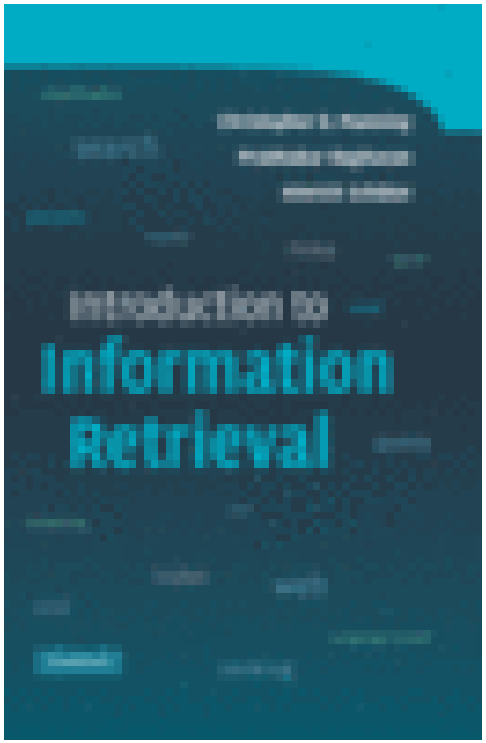


O'Connor, et al. In ICWSM-2010

Textbook

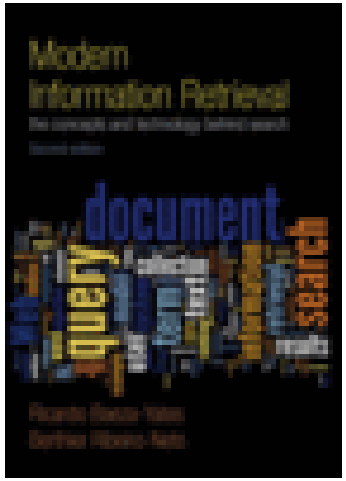
- Introduction to Information Retrieval
- Free online version

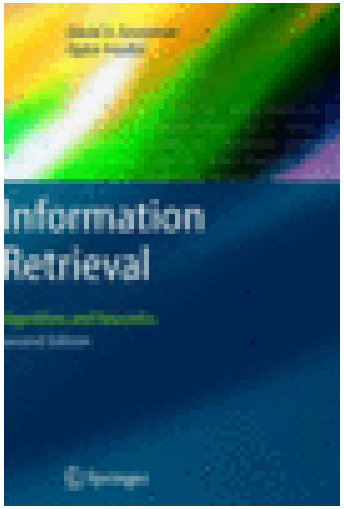
http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf



References

- Search Engines: Information Retrieval in Practice
- Modern Information Retrieval
- Information Retrieval: Algorithms and Heuristics
- Mining the Web: Discovering Knowledge from Hypertext Data





Tentative topics

- Overview (Background, history, basic concepts)
- Text processing (Zipf's Law, tokenizing, stemming, indexing)
- Query operations (Query expansion, structural queries, relevance feedback)
- Classic Retrieval models (Boolean model, vector space model, TF-IDF weighting, text-similarity metrics)
- Statistical language models
- PageRank
- Recommendation systems (Collaborative filtering, content-based filtering, implicit user feedback)
- Text clustering and categorization
- Big data with Hadoop

The goal

- Learn the techniques behind Web search engines and recommendation systems
- Get hands-on project experience by building information retrieval systems
- Lead to the amazing job opportunities in Search Industry and E-commerce companies such as Google, Bing, Yahoo!, Amazon, etc.
- Lay a foundation to do cutting-edge research

Prerequisites

- AMTH 108 (Probability and Stat)
- MATH 53 (Linear Algebra)
- Proficient in one programming language

Assignments

- Written assignments

- Projects
  - Search engines
  - Recommendation systems
  - Hadoop

Exams

- Midterm
- Final

Based on lecture contents and assignments

Grading Policy

- Written assignment: 10%
- Projects: 40% (10%+20%+10%)
- Midterm exam: 20%
- Final exam: 30%

- ➢ Late submission will be penalized 10% per day (with weekends counting as one day)
- ➢ The assignments must be done individually
- ➢ It is safe to start early…

Course information

- Instructor: Yi Fang
- Class meets TTh, 12:10-1:50pm
- Course materials on Camino
- Contact info: EC 246, yfang@scu.edu
- Homepage: http://www.cse.scu.edu/~yfang
- Office hours
  - Tuesday 2-3pm, Friday 1-2pm
  - By appointment
  - Anytime by email

Conclusion

- Overview of Web Information Management
- Course policy
- Next lecture: basic and core concepts