

# COEN 169

---

## Web Information Management

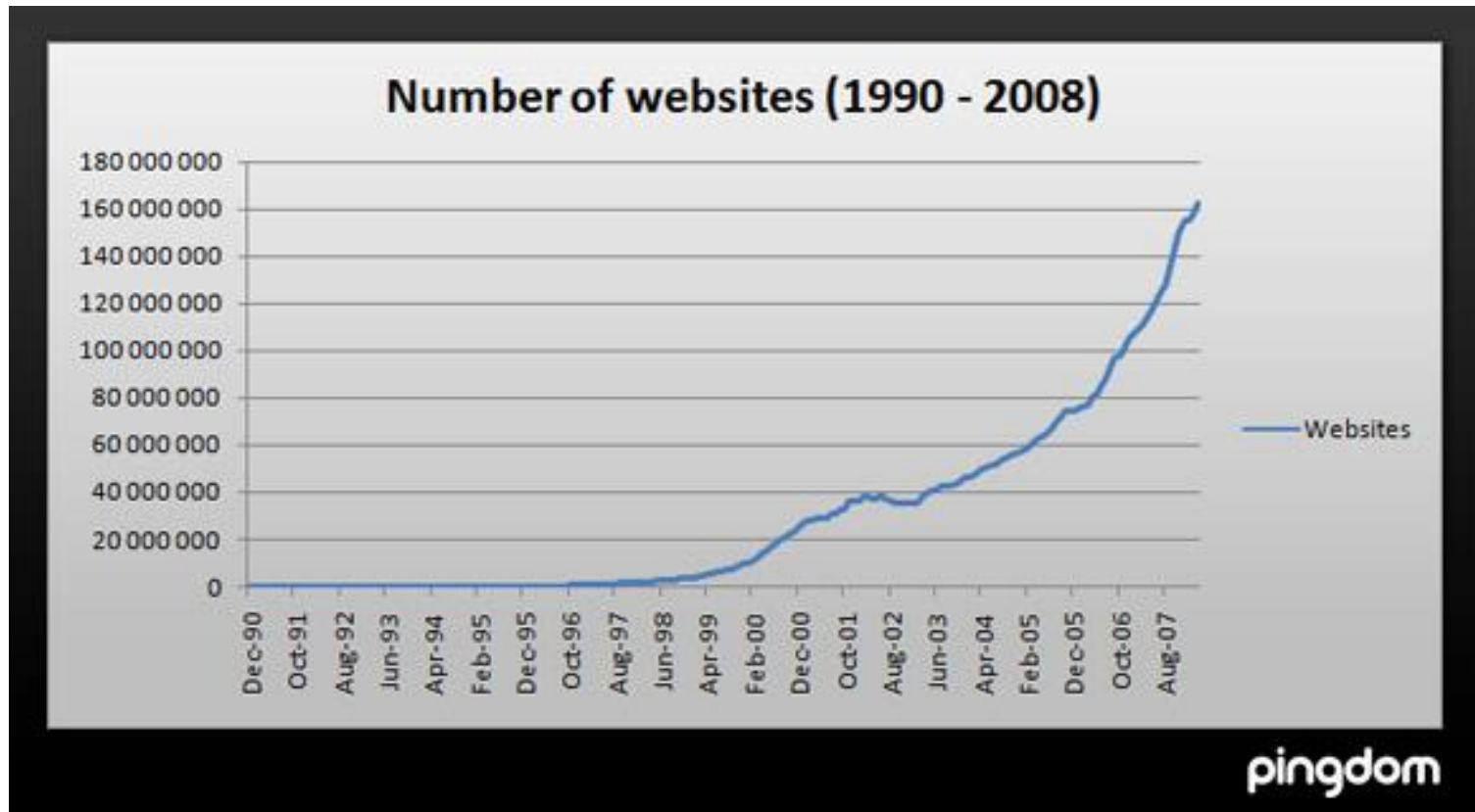
Yi Fang, Ph.D.

<http://www.cse.scu.edu/~yfang>

Department of Computer Engineering

Santa Clara University

# Web growth



# Web Information Management

---

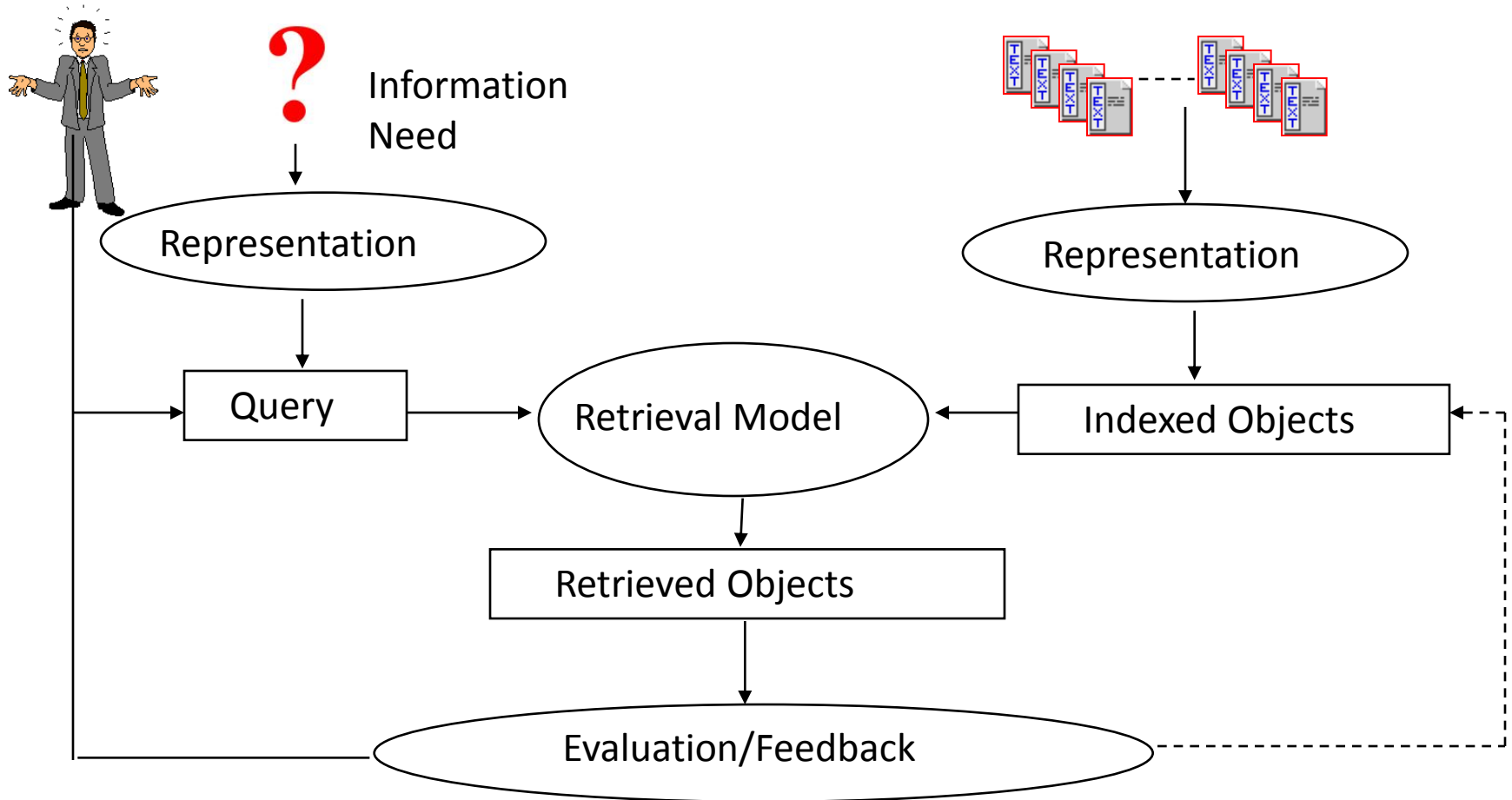
Theory, design, and implementation of information systems that process, organize, analyze large-scale information on the Web

- Search engines
- Recommendation systems

# History of Search Engines

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

# 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

---

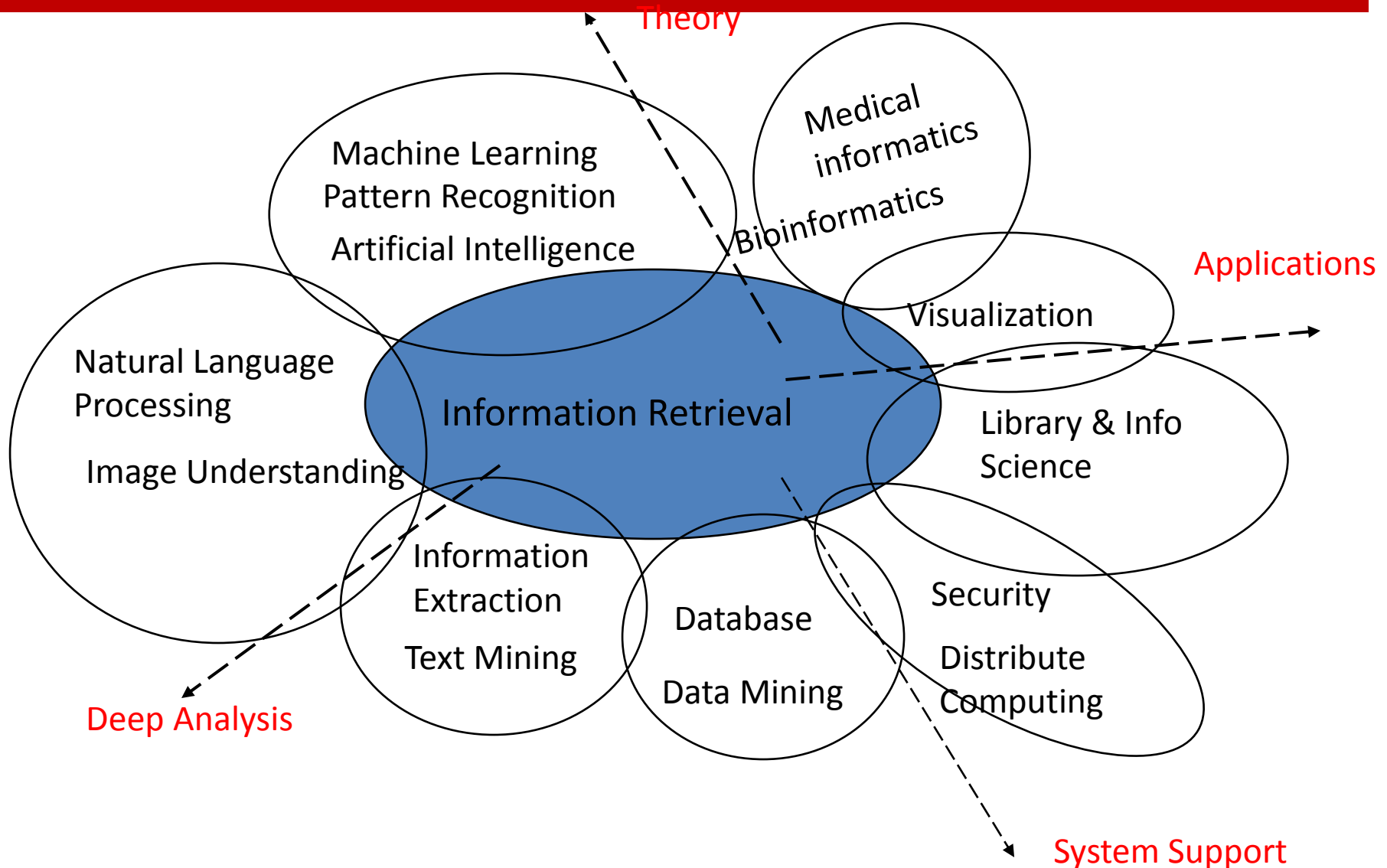
## Relational Database Management Systems (RDBMS):

- 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

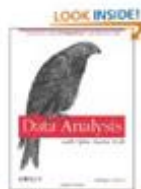
- **Amazon.com** recommends books based on your purchase history (and others')

Yi, Welcome to Your Amazon.com (If you're not Yi Fang, click here.)

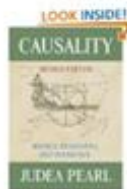
## Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#).

Page 2 of 35 (Start over)



[Data Analysis with Open Source Tools](#)  
(Paperback) by Philipp K. Janert  
★★★★☆ (23) \$25.30  
[Fix this recommendation](#)



[Causality: Models, Reasoning and Inference](#)  
(Hardcover) by Judea Pearl  
★★★★☆ (5) \$42.79  
[Fix this recommendation](#)



[Design Patterns: Elements of Reusable Object-Oriented Software](#)  
(Hardcover) by Erich Gamma  
★★★★☆ (297) \$32.99  
[Fix this recommendation](#)



[Code Complete: A Practical Handbook of Software Construction](#)  
(Paperback) by Steve McConnell  
★★★★☆ (150) \$29.54  
[Fix this recommendation](#)



[Fundamentals of Embedded Software: Where C and Assembly Meet](#)  
by Daniel W. Lewis  
★★★★☆ (11) \$82.96  
[Fix this recommendation](#)

# Movie Recommendation systems

- **Netflix** predicts other “Movies You’ll Love”



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

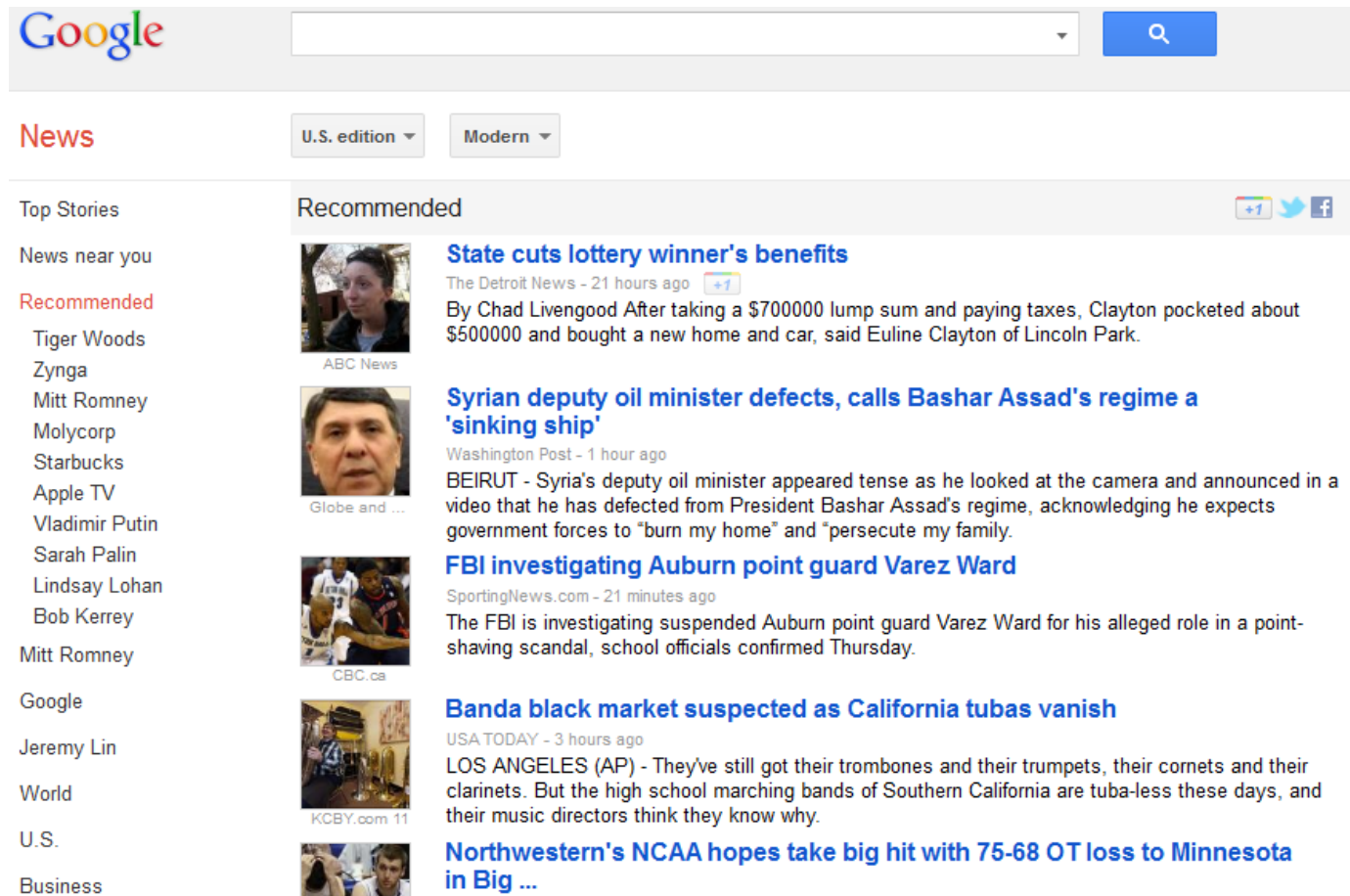
# Recommendation algorithms



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

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

# News Recommendation



The screenshot shows the Google News homepage. At the top is the Google logo and a search bar. Below the logo, there are tabs for 'News', 'U.S. edition', and 'Modern'. The main content area is divided into two sections: 'Top Stories' and 'Recommended'. The 'Recommended' section is highlighted with a blue header and contains four news articles. Each article has a thumbnail image, a headline, a source, and a timestamp. The articles are: 1. 'State cuts lottery winner's benefits' by The Detroit News, 21 hours ago. 2. 'Syrian deputy oil minister defects, calls Bashar Assad's regime a 'sinking ship'' by Washington Post, 1 hour ago. 3. 'FBI investigating Auburn point guard Varez Ward' by SportingNews.com, 21 minutes ago. 4. 'Banda black market suspected as California tubas vanish' by USA TODAY, 3 hours ago. The sidebar on the left lists 'Top Stories' and 'News near you' with a list of names and topics: Tiger Woods, Zynga, Mitt Romney, Molycorp, Starbucks, Apple TV, Vladimir Putin, Sarah Palin, Lindsay Lohan, Bob Kerrey, Mitt Romney, Google, Jeremy Lin, World, U.S., and Business.

Google

News

U.S. edition Modern

Top Stories

News near you

Recommended

Tiger Woods

Zynga

Mitt Romney

Molycorp

Starbucks

Apple TV

Vladimir Putin

Sarah Palin

Lindsay Lohan

Bob Kerrey

Mitt Romney

Google

Jeremy Lin

World

U.S.

Business

Recommended

**State cuts lottery winner's benefits**

The Detroit News - 21 hours ago

By Chad Livengood After taking a \$700000 lump sum and paying taxes, Clayton pocketed about \$500000 and bought a new home and car, said Euline Clayton of Lincoln Park.

**Syrian deputy oil minister defects, calls Bashar Assad's regime a 'sinking ship'**

Washington Post - 1 hour ago

BEIRUT - Syria's deputy oil minister appeared tense as he looked at the camera and announced in a video that he has defected from President Bashar Assad's regime, acknowledging he expects government forces to "burn my home" and "persecute my family."

**FBI investigating Auburn point guard Varez Ward**

SportingNews.com - 21 minutes ago

The FBI is investigating suspended Auburn point guard Varez Ward for his alleged role in a point-shaving scandal, school officials confirmed Thursday.

**Banda black market suspected as California tubas vanish**

USA TODAY - 3 hours ago

LOS ANGELES (AP) - They've still got their trombones and their trumpets, their cornets and their clarinets. But the high school marching bands of Southern California are tuba-less these days, and their music directors think they know why.

**Northwestern's NCAA hopes take big hit with 75-68 OT loss to Minnesota in Big ...**

➤ **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

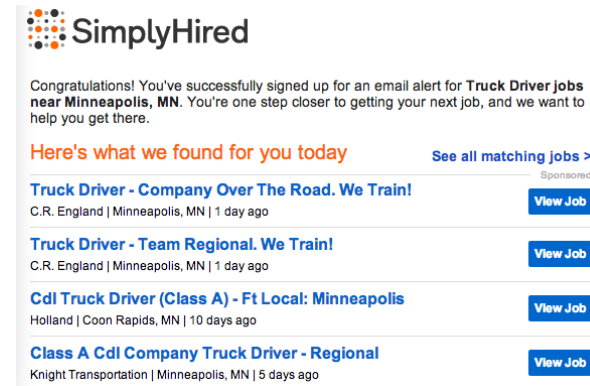


Figure 1: An example of *Simply Hired*'s email alert service for job recommendation.



# Point-of-Interest Recommendation



---

- Foursquare check-in
- Google Place API





# IR Applications: Text Categorization



**News**

U.S. edition ▼Modern ▼

Jeremy Lin

World

U.S.

Business

Elections

Technology

iCloud

Mobile Technology

Mobile Industry

Charlie Miller

Go Daddy

Online Security

Apple


EBay

Search Engines

Instagram

Entertainment

Search Engines


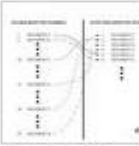




Search Engine Land


**Guide to Understanding the Searcher Experience**

Search Engine Land - 5 hours ago

communicate aboutness well enough when search engines are not yet able to clearly determine aboutness from actual document content, such as a graphic image (GIF, JPEG, PNG)?



Search En...Business 2...Business 2...WorldDenta...




The Atlantic

**What to Make of Google's Decision to Block the 'Innocence of Muslims' Movie**

The Atlantic - 2 hours ago

The attacks on U.S. missions abroad this week have been a test for Google's "bias in favor of free expression." RTR37VPT-615.




CITY LIMITansas City

**Google Fiber Issues Public Challenge: Get Up To Speed!**

TIME - 6 hours ago

In addition to building the world's largest Internet search engine, Google was furiously buying up so-called "dark fiber" the unused long-haul underground cable to A document for the data center search.

# IR Applications: Document Clustering



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 [context](#). Sun Microsystems currently maintains and updates Java n

# IR Applications: Information Extraction

OPUS International, Inc., an executive search firm focusing on the Food Science industry. - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop

Address [http://www.foodscience.com/jobs\\_midwest.html#top](http://www.foodscience.com/jobs_midwest.html#top)

Links AMEX Rewards Time DogHouse My Yahoo!

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

OPUS INTERNATIONAL INC.

About | Staff | Job

food sci

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

**Test Kitchen-  
Consumer Food Relations**

Major food manufacturer in Chicago area seeks a consumer food professional to write all recipes. Will make presentations; will be a key player in a cross-functional team. Requires a BS in human ecology, nutrition, Food Science, or related field. A minimum three years' experience.

Contact: Moira: e-mail  
1-800-488-2611

**Ice Cream Guru**

If you dream of cold creamy chocolate or gooey, gooey cookie, there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream brand. Will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.

Contact: Susan: e-mail  
1-800-488-2611

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



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/](http://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](http://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](http://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](http://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](http://en.wikipedia.org/wiki/West_Coast_Conference)

Jump to [Former members](#): **University** of the Pacific (1952–1971) (now a **member** of

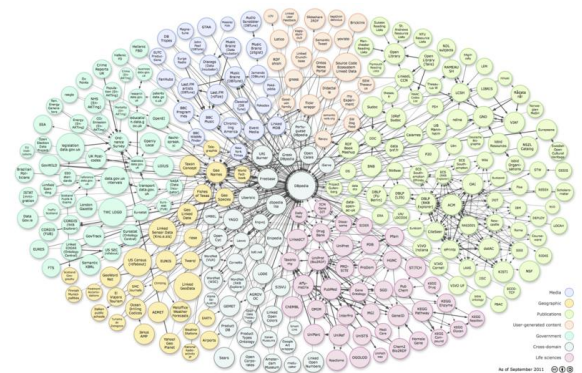
# Why Entity Search?

## ➤ 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](http://www.indure.org)



The screenshot shows the homepage of the INDURE website. The header features the 'indure' logo with a stylized orange 'i' and the text 'INDIANA Accelerate Your Business'. Below the logo is the tagline 'INDIANA DATABASE FOR UNIVERSITY RESEARCH EXPERTISE'. To the right of the logo is a dropdown menu labeled 'Filter By Institution'. Below the header is a navigation bar with links: 'Home', 'Advanced Search', 'Research Areas', and 'Faculty and Admin, login here'. The main content area includes a photograph of a woman in a lab coat working in a laboratory. To the right of the photo is the text 'Welcome to INDURE Indiana Database of University Research Expertise' and 'Version 1.0 August 15th, 2008'. Below this is a search bar with a 'Search' button.

**indure** | **INDIANA**  
INDIANA DATABASE FOR UNIVERSITY RESEARCH EXPERTISE | Accelerate Your Business

Filter By Institution

Faculty and Admin, login here

Home Advanced Search Research Areas Faculty and Admin, login here


**Welcome to INDURE**  
**Indiana Database of University Research Expertise**

Version 1.0 August 15th, 2008


Search

you may use INDURE to search for

# Question Answering



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](http://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](http://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](http://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


**Linked in** Account Type: Basic | [Upgrade](#)

Yi Fang ▾ [Add Connections](#)



Home Profile Contacts Groups Jobs Inbox Companies News More


People ▾ Search...  Advanced

[What is self-plagiarism? - Download a free paper to understand self-plagiarism and how to avoid it.](#)





Share an update


 Attach a link  [Share](#)





**LinkedIn Today** recommends this news for you


**Top News: Facebook's Backtrack, Who's**  
[linkedin.com](#)

**Toyota Hired 6 Ad Agencies To Create This New 3-Word**  
[businessinsider.com](#)


**Intel Confirms Freefall of Server Giants HP, Dell, and**  
[wired.com](#)


**Yangqiu Song**, Associate Researcher at Microsoft  
[Connect](#)


**Qin Iris Wang**, Research Scientist at TheFind, Inc.  
[Connect](#)


**forrest yang**, Software Engineer at Google  
[Connect](#)

[See more »](#)

 Ads by LinkedIn Members

**Chromebooks for Education**  
Check out the new Chromebook. Efficient to manage, easy to use!

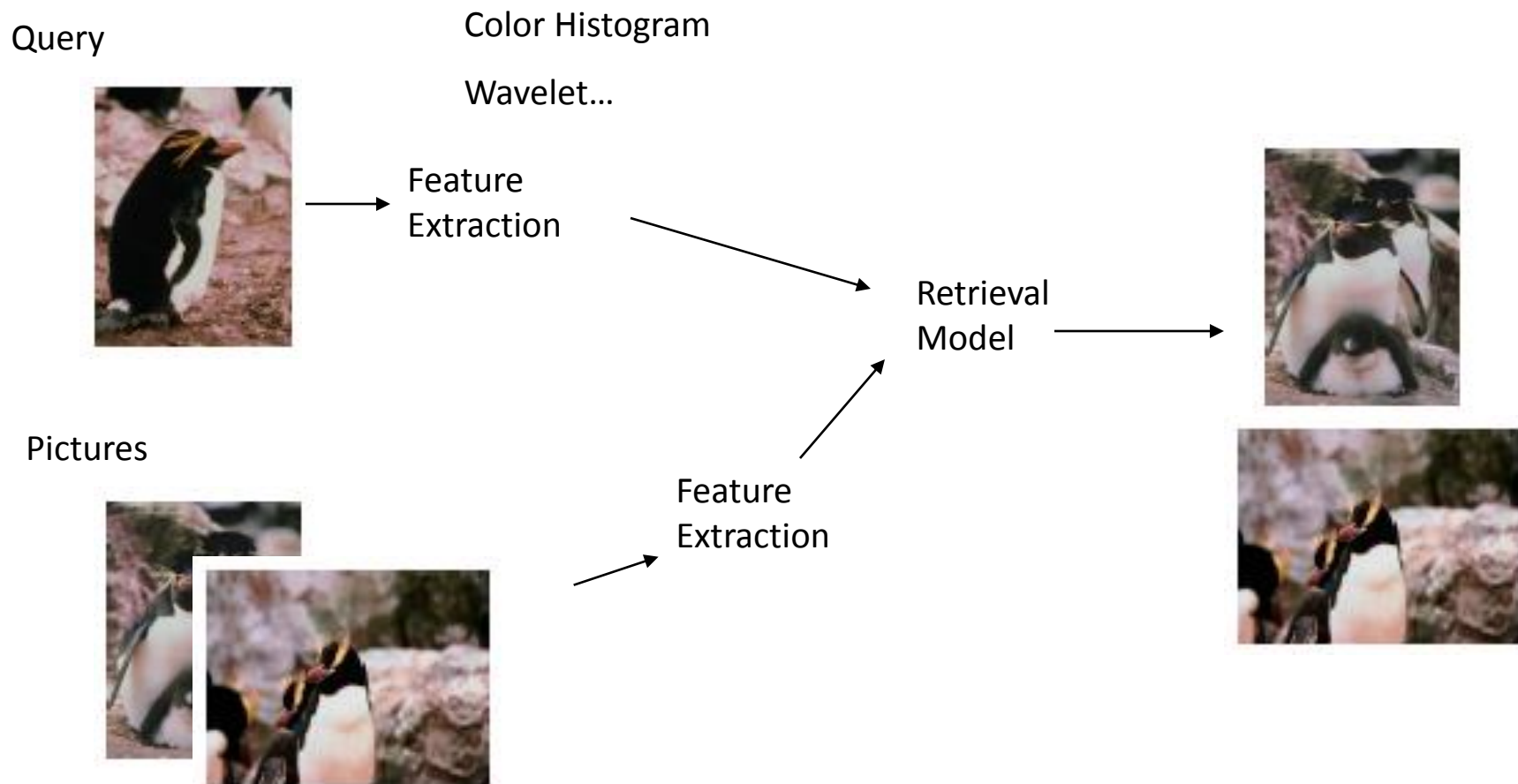
**Change Healthcare Future**  
Earn MS/PhD in Mind-Body Medicine- Online courses. Apply now for Spring 2013

**Top Ranked MBA | Purdue**  
Top 30 Ranked MBA Degrees Designed for Working Professionals. Get Info:

Steve Qin is now connected to [Chengzhi Wang](#), — and [Bingfeng Li](#), SDET at

# IR Applications: Multimedia Retrieval

---



# Detecting Road Hazards Using Twitter

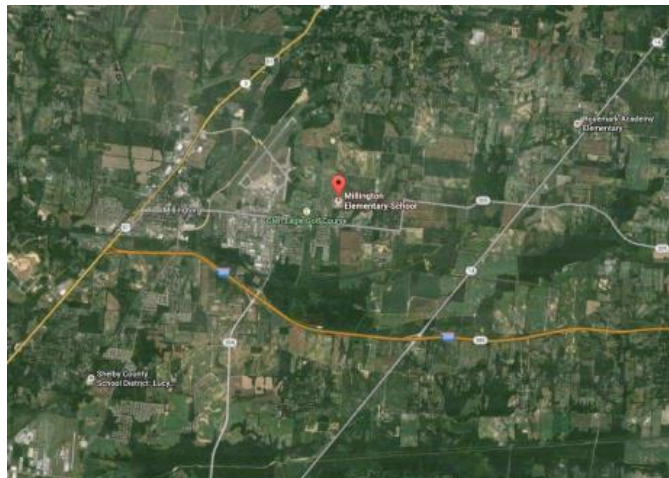
---



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](https://twitpic.com/aawx1)



# Trending topic discovery

The screenshot shows the Yahoo! homepage interface. At the top, there's a navigation bar with links for Web, Images, Video, Local, Apps, and More. Below this is the Yahoo! logo and a search bar. The date "Saturday, March 10, 2012" is displayed. On the left, there's a "YAHOO! SITES" menu with links to Autos, Dating, Finance (Dow), Flickr, Games, and Horoscopes. The main content area features three images of NBA players: Biggie Smalls, a woman, and Kirk Cameron. Below these images is the text "NBA stars who could be on the move". To the right, a red-bordered box highlights the "TRENDING NOW" section, which lists ten topics in two columns. The topics are: 01 Biggie Smalls, 02 Mystic Pizza owners, 03 Whitney Houston will, 04 Kirk Cameron, 05 Oil prices, 06 Ellen DeGeneres, 07 Goldie Hawn, 08 Coke and Pepsi cha..., 09 Tax deductions, and 10 Daylight saving time. At the bottom right, there's an "AdChoices" link.

Make Y! your homepage

Web Images Video Local Apps More

YAHOO!

Saturday, March 10, 2012

HI, Y! Sign Out

14 MAIL 14 new emails

YAHOO! SITES

- Autos
- Dating
- Finance (Dow)
- Flickr
- Games
- Horoscopes

NBA stars who could be on the move

**TRENDING NOW**

01 Biggie Smalls	06 Ellen DeGeneres
02 Mystic Pizza owners	07 Goldie Hawn
03 Whitney Houston will	08 Coke and Pepsi cha...
04 Kirk Cameron	09 Tax deductions
05 Oil prices	10 Daylight saving time

AdChoices

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



# Sentiment Analysis

---

Positive or negative movie review?



- unbelievably disappointing



- Full of zany characters and richly applied satire, and some great plot twists



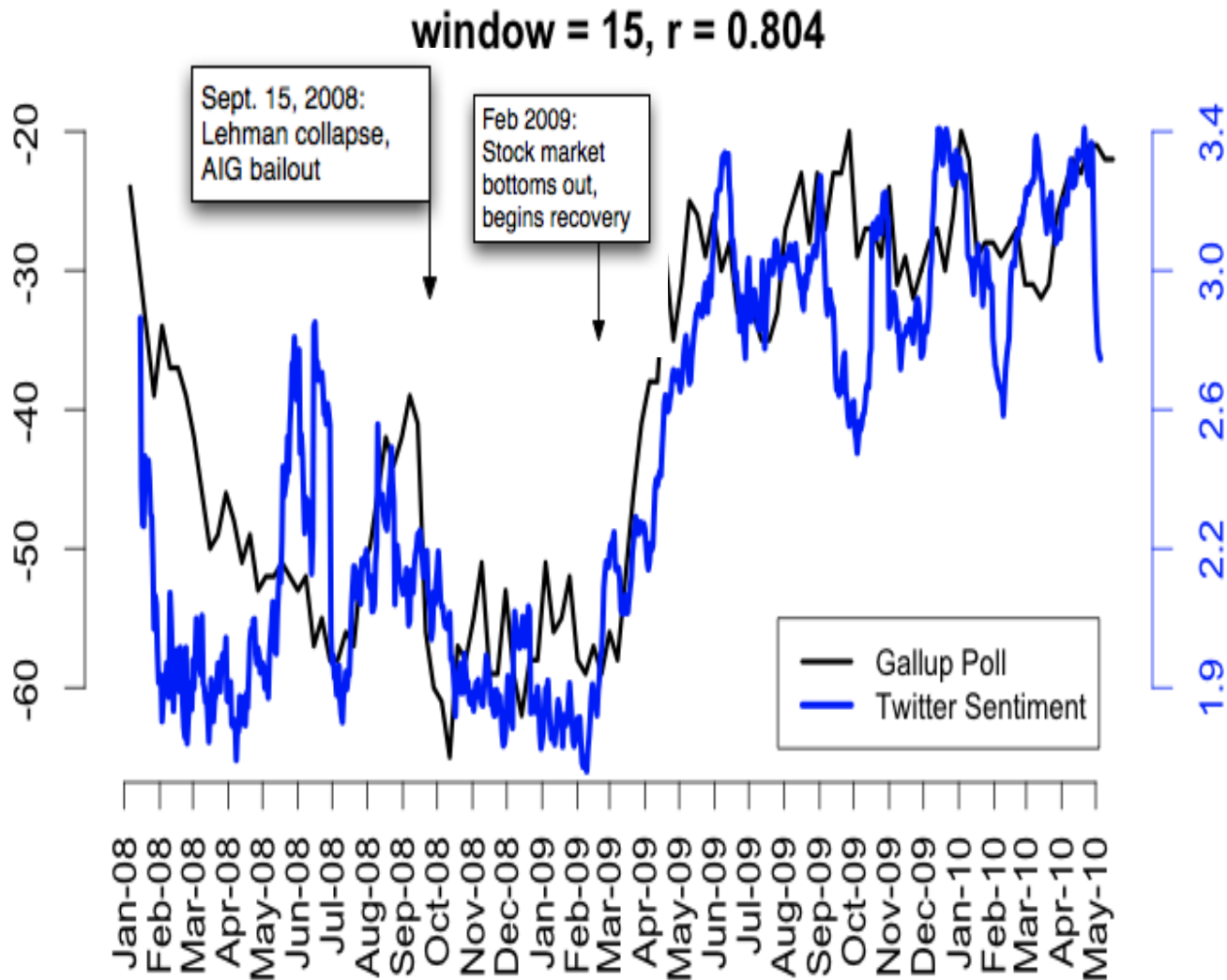
- this is the greatest screwball comedy ever filmed



- It was pathetic. The worst part about it was the boxing scenes.



# Twitter sentiment versus Gallup Poll of Consumer Confidence



# Textbook

---

- Introduction to Information Retrieval
- Free online version

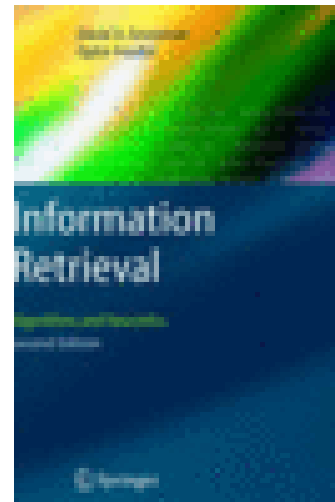
<http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>



# 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](mailto: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