Introduction to Search/Retrieval Technologies

Yi Zhang
Information System and Technology Management
IRKM Lab
University of California Santa Cruz
Information Retrieval and Knowledge Management Lab

- **Focus:** information retrieval, data mining, applied machine learning
- **Director:** Professor Yi Zhang
- **Staff:** James Snook (Research Software Engineer) Tao Hong (Research Scientist)
- **Graduate Students:**
  - Aaron Michelony (Ph.D.) Sarah Tyler (Ph.D.)
  - Yunfei Chen (Ph.D.) Lanbo Zhang (Ph.D.)
  - Jian Wang (Ph.D.) Shawn Wolf (NASA, Ph.D.)
  - Qi Zhao (Ph.D.)
- **Undergraduate researchers:**
  - Gabriel Sanchez; Benjamin Chow, Tiffany Yee, Leland Takamine, etc.
- **Alumni placement:** Facebook, LinkedIn, Google, etc. various start up companies, graduate schools
Which Search Engine(s) Do You Use?
What is Traditional Information Retrieval?

- A collection of unstructured or semi-structured documents
- User inputs a query to express his/her information need
- A retrieval system returns a list of documents to satisfy the information need of the user
Context of Search

Models
- Statistics, optimization
- Artificial Intelligence
  - Machine learning
  - Natural language processing
- Human computer interaction
- Computer networks

Applications
- Bioinformatics
- Business applications
- Medical informatics
- Digital libraries
- Web search
- Online marketing
- Online advertising

Systems
- Database
- Computer systems
- Security
Major Technologies: Go Beyond Traditional Search

Access Information

- Categorization
- Filter
- Search

Clustering

Information Retrieval

Generate Knowledge

- Question and Answer
- Information Extraction
- Text mining

Text (possible meta data)
Outline

- Introduction
- History of search
- Web search
- Future of search
  - Our research focus
History of Search/Retrieval

- 3rd Century B.C. Library of Alexandrian
  - Catalogs and classifications (controlled vocabulary)
  - Alphabetization
- 1247 First Concordance of the Bible
  - Invention of the inverted list data structure
History Continues

- 1930’s Punch Card
  - Manual retrieval system
  - Satisfy Boolean query
  - Card: keyword
  - Document
  
  Retrieval algorithm:
  
  The documents corresponding to the position where light falls through all “query” cards are the wanted documents.
History Continues

- 1947 Vannevar Bush’s Memex
  - Some ideas are part of our life
    - theoretical proto-hypertext
  - Some 1947 ideas may be the future
    - Vision of personalized information management system
History Continues

- 1960’s and 1970’s Computer based IR
  - Quantitative aspects of text and the models that were proposed were based on word frequencies and word occurrences
  - Small scale: library
- 1994 Web Search Engine
- 1997 Image and video retrieval
- 1999 Question and answering

Based on Bruce Croft and Ned Fielden
A Typical/Simple Retrieval Process

User’s Information Need (GUI, user models)

Representation
(stop, stem, NLP, meta data, expansion, structure, phrase)

Query (query language)

Indexed Objects (inverted index)

Comparison (retrieval models)

Evaluation/User Feedback

Retrieved Documents
Issues in Retrieval

- How to represent text
- How to represent the information needs of the user
- How to compare representations (rank documents)
- How to evaluate the effectiveness of retrieval
Different Retrieval Models

- Boolean model: query as rules for identifying relevant documents
- Vector space model: query as a small relevant document
- Bayesian inference model: query as an expression of the information need
- Language models: query as a sample of the relevant documents
How to evaluate the effectiveness of retrieval

- Evaluation
  - Measures the ability of the system to find relevant document

- What is “relevance”?
  - Difficulty to define
  - "usefulness", "related", ...

- User judgments are used for evaluation
  - People disagree on what is relevant (20%)
  - Same person isn’t consistent
  - Judgments depends on the context
Web Search: Using the Web Structure
Basic Web Search Engine Infrastructure

The Web

- Crawling
- Indexing
- Archive
- Index

Query Processor

query

Search
Link Based Ranking

- Intuition: link is similar to “citation” in the literature. Better quality pages get more citation
  - Example: www.soe.ucsc.edu ...

- Most web search engines use link based ranking
  - Google’s PageRanking: a page is important if a lot of other pages link to it. It is especially important if other important pages link to it.
Google’s PageRanking

- Random walk model

- Start at a random page
- At each step, go out of the current page along one of the links on that page with equal probability
- Each page has a long term visit rate in the “steady state”: use this rate as the page ranking score
PageRank: Variations

- Topic specific Pagerank (Have02)
- Personalized Pagerank
  - Using non-uniform teleportation
Discussion: SEO

- If you want to boost your pages in a search engine only based on PageRank, what will you do?
  - If you are a spammer?

- If you run the search engine, what kind of technology will you use to fight against the spammers.
Key Research Directions

- Natural Language Processing
- Personalization & Proactive retrieval: recommendation systems
- Social search
What is NLP?

Arabic text

يَجِبُ عَلَى الإنسانِ أن يَكُونَ أمينًا وصادقاً مَعَ نَفْسِه وَمَعَ أَهْلِه وَجِيْرَانِه وَأَن يَبْذُلْ ...

... كَلَ جُهَدٍ في إِعْلاءِ شَأْنِ الوَطَنِ وَأَن يَعْمَلَ عَلَى مَا

How can a computer make **sense** out of this **string**?

**Morphology**
- What are the basic units of meaning (words)?
- What is the meaning of each word?

**Syntax**
- How are words related with each other?

**Semantics**
- What is the “combined meaning” of words?

**Pragmatics**
- What is the “meta-meaning”?

**Discourse**
- Handling a large chunk of text

**Inference**
- Making sense of everything

From Chengxiang Zhai
An Example of NLP

A dog is chasing a boy on the playground

- **Det**: Dog(d1), Boy(b1), Playground(p1), Chasing(d1,b1,p1).
- **Noun**: Dog, Boy, Playground.
- **Verb**: Chasing.
- **Prep**: on, the.

**Semantic analysis**

- Scared(x) if Chasing(_,x,__).
- Scared(b1)

**Lexical analysis (part-of-speech tagging)**

- *Dog(d1)*
- *Boy(b1)*
- *Playground(p1)*
- *Chasing(d1,b1,p1)*

**Syntactic analysis (Parsing)**

- Sentence
- Verb Phrase
- Prep Phrase
- Noun Phrase

**Pragmatic analysis (speech act)**

- A person saying this may be reminding another person to get the dog back…

From Chengxiang Zhai
What Machine Can do Now?

The State of the Art

A dog is chasing a boy on the playground

Det Noun Aux Verb Det Noun Prep Det Noun

Noun Phrase Complex Verb Noun Phrase

Verb Phrase Verb Phrase

Sentence Prep Phrase

Semantics: some aspects
- Entity/relation extraction
- Word sense disambiguation
- Anaphora resolution

Inference: ???

POS Tagging: 97%

Parsing: partial >90%?

Speech act analysis: ???

From Chengxiang Zhai
Personalized Proactive Information Retrieval has Many Potential Markets

- You need to know new products related to yours
  - E.g., from the literature, web, or discussion groups, news
- You want to know good restaurants in Santa Cruz
- You want good movies for your family

![Amazon.com](amazon.com)

![buy.com](buy.com)

![Netaflix](netflix.com)

![Circuit City](circuitcity.com)

![Ask.com](ask.com)

![eBay](eBay.com)

![Yahoo Shopping](shopping.yahoo.com)
Stoneridge Reports Second-Quarter 2006 Results

- Net Income Increase of 74% from Prior Year
- Company Raises Full-Year 2006 Earnings Outlook

WARREN, Ohio, July 28 /PRNewswire-FirstCall/ -- Stoneridge, Inc. (NYSE: SRI) today announced net sales of $185.5 million and net income of $4.9 million, or $0.21 per diluted share, for the second quarter ended July 1, 2006.

Net sales increased $5.2 million, or 2.9 percent, to $185.5 million, compared with $180.3 million for the second quarter of 2005. The increase in sales was primarily due to strong demand in the Company’s commercial vehicle markets. The effect of foreign currency translation reduced second-quarter net sales by approximately $0.9 million compared with the same period in 2005.

Net income for the second quarter was $4.9 million, or $0.21 per diluted share, compared with net income of $2.8 million, or $0.12 per diluted share, in the second quarter of 2005. The increase in net income was primarily attributable to improved
Potential Terrorist Alert?
Even if You Do Not Work…

Overwhelmed by information ….
Today’s Search Engines Can Help But … Not Enough!

- Search engine focus: user pulls information from the system using a query
  - Short term information need (ad hoc search)

- The task: proactive system pushes information to the user without requiring any explicit user query
Proactive Personalized Search

Adaptive Filtering System
(binary classifier)
(user profile)

item stream

Recommended items

feedback

accumulated items

initialization

free text query

learning user profile

32
How Do you Know About the User Preferences?

- Task: will this user like the item (document, movie, product, etc.)?
"There is an awful lot you can tell about a person by their shoes .. Where they're going, where they've been."

Forrest Gump, Tom Hanks
Find Your Soul Mate in Less than 10 Minutes

You can tell a lot about people by their reactions and opinions about other people.

-8 Minute Dating
FBI can read!....your mind!

According to leading FBI agent, Joseph Navarro, you can tell alot about people by doing two simple things:

1. Read their body language
2. Learn when expressions are "faked"
Major Approaches

- Approach 1: collaborative filtering
- Approach 2: content based adaptive filtering
- The combination of both
Build Proactive Information Retrieval System with Desirable Characteristics

What can a software with human intelligence do? (desirable characteristics)

- Use heuristics
- Ask good questions
- Use context and implicit feedback
- Learn from other users/Social networks

Our solution for a computer

Unified Framework

- Bayesian Prior
- Bayesian Active Learning
- Graphical Models
- Bayesian Hierarchical Models
Video Demo Built by Students
Mobile Demo: Personalized Videos

User Feedback

- Get user feedback on system recommendations
- Re estimate user model based on the new feedback
- Predict items that reflect user’s taste as portrayed by his feedback
Performance

• News recommendation

• Netflix movie recommendation: RMSE = 0.8874 (6.73% better than Netflix’s own system)

• Epinion.com products recommendation: MSE = 0.49 on cases that we can use trust network between users
Proactive Retrieval for Classrooms

- Users: students who are reading an article and stumbled over a word they don’t understand
- Our solution: multimedia multilingual context sensitive search engine to find online resources that helps the students better understand the word tecwave.org

- Demo

- Initial collaborators
  - US Department of Education
  - Shoreline Middle School in the Live Oak school district
  - Santa Clara Unified School District
Why It’s an Exciting Research Topic?

- Opportunities: the problem is well motivated with many potential applications
- Intellectual Merit: the problem is challenging and needs research
  - Small, noisy, biased data for the machine to learn
  - Whole interactive learning process is extremely hard to model
  - Computation complexity
  - Factors influence user decision is multi faceted
- Broader impact: the solution will has a broader impact in other areas such as artificial intelligence, marketing..
Current Research Projects

- Proactive Personalized Information Retrieval
  - Recommender systems
  - Search personalization
  - Personalized internet advertising and marketing

- TecWave: Teaching with Computers: Word Annotation for Vocabulary Education

- Adaptive Information Filtering
  - Content based information filter for analysts
  - Content based intrusion detection

- Distributed Peta Scale Semi Structured Data Search with Facets

- Applied Machine Learning
  - Transfer learning, active learning, feature learning
Sponsors of IRKM Lab

- Air Force Office of Scientific Research
- Department of Education
- Nokia
- NEC
- Bosch
- Los Alamos National Laboratory
- Google
- Yahoo!
- Microsoft Research
What I Teach

- ISM260 Information Retrieval (Fall)
- ISM58 System Analysis and Design (Winter)
- ISM245 Data Mining (Winter)
- ISM280I Information Retrieval and Knowledge Management Seminar (Fall)
Interested?

- Undergraduate research opportunities in summer
  - To be sponsored by National Science Foundation
- TREC competition/evaluation
  - Blog, Enterprise, Legal data
  - Genome data, Question and Answering, Video retrieval, Spam filtering