Introduction to Search and Recommendation

Yi Zhang
Associate Professor
Information Retrieval and Knowledge Management Lab
Graduate Director, Technology and Information Management
University of California Santa Cruz
Information Retrieval and Knowledge Management Lab

- Focus: information retrieval, intelligent data and big analytics, economics for computation
- **Director:** Professor Yi Zhang
- 8 Ph.D. students and 3 M.S. students
- Undergraduate researchers
- Alumni placement: Facebook, LinkedIn, Google, etc. various start up companies, graduate schools
- Sponsors of lab research:

![Sponsors Logos]
Which Search Engine(s) Do You Use?
Context of Search

- Models
  - Statistics, optimization
  - Artificial Intelligence
    - Machine learning
    - Natural language processing
  - Human computer interaction
- Systems
  - Computer networks
- Applications
  - Bioinformatics
  - Business applications
  - Medical informatics
  - Digital libraries
  - Web search
  - Online marketing
  - Online advertising
  - Database
  - Computer systems
  - Security
  - ...
Major Technologies: Go Beyond Traditional Search

- Access Information
- Generate Knowledge

- Categorization
- Clustering
- Question and Answer
- Information Retrieval
- Information Extraction
- Text mining
- Filter
- Search

Text (possible meta data)
Outline

- Introduction
- History of search
- Web search
- Proactive search: recommendation systems (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
An Example

Query: Michael Jordan
What’s my information needs?
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
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
Key Research Directions

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

- What are the basic units of meaning (words)?
- What is the meaning of each word?
- How are words related with each other?
- What is the “combined meaning” of words?
- What is the “meta-meaning”?
- Handling a large chunk of text
- Making sense of everything

From Chengxiang Zhai
An Example of NLP

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

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

  + Scared(x) if Chasing(_,x,___).

  - Scared(b1)

- **Syntactic analysis (Parsing)**

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

- **Pragmatic analysis (speech act)**

  - From Chengxiang Zhai
What Machine Can do Now?
The State of the Art

A dog is chasing a boy on the playground

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

Inference: ???

From Chengxiang Zhai
Web Search: Using the Web Structure
Retrieval results
Basic Web Search Engine Infrastructure

- The Web
- Archive
- Index
- Indexer
- Indexing
- Query Processor
- crawling
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.
Different Retrieval Models

- Boolean model: Rules for Identifying relevant documents
- Vector space model: a small relevant document
- Bayesian inference model: an expression of the information need
- Language models: sample of the relevant documents
- Link based models (web)
- Learning to rank
Learning to Rank: Problem Setting

- Given: user query $q$, web pages $D=\{d_1, d_2, \ldots, d_N\}$
- Each web page $d_i$ is represented as a feature vector
  - Text match score with title, anchor text, headings, bold text, body text, . . . , of $d_i$ as a hypertext document
  - Pagerank, topic-specific Pageranks, personalized Pageranks of $d_i$ as a node in the Web graph
  - Estimated location of user, commercial intent, . . .
- Goal: a single scoring function on $(q, D)$ so as to induce a ranking on $D=\{d_1, d_2, \ldots, d_N\}$?
### Ranking Features for (Document, Query) Pairs

<table>
<thead>
<tr>
<th>Local</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Length</td>
<td>Popularity/Pagerank</td>
</tr>
<tr>
<td>Page Quality</td>
<td>Web Connectivity</td>
</tr>
<tr>
<td>...</td>
<td>Document Clusters</td>
</tr>
<tr>
<td>Term frequency</td>
<td>Anchor-text</td>
</tr>
<tr>
<td>Proximity</td>
<td>Number of results</td>
</tr>
<tr>
<td>Section matches</td>
<td>Result Set Clusters</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>Independent</th>
<th>Dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Length</td>
<td>Page Quality</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

From Ahmed
Learning to Rank Approaches

- **Pointwise learning**
  - **Input:** single document  
    **Output:** relevance score
  - As a traditional classification/regression problem
  - Loss function defined over each document (Evaluation: difference between the score assigned by the algorithm and the “true score”)

- **Pairwise learning**
  - **Input:** document pairs \((d_i, d_j)\)  
    **Output:** partial order preference
  - As a special classification/regression problem to classify whether \(d_i\) is prefered to \(d_j\) or not.
  - Loss function defined over each document pair (Evaluation: count the number of reversed pairs)

- **Listwise learning**
  - **Input:** document collections  
    **Output:** ranked document list \(y\)
  - Lost function defined over each ranking list
Proactive Search/Retrieval: Recommendation systems
Proactive Information Retrieval for Stock

Proactively recommend news/blog/message related to your stock portfolio
Proactive Information Retrieval for Job
Proactive Information Retrieval Everywhere

- Want to find a good restaurant in Santa Cruz
  - Based on your health condition, taste preferences, current location and time, social context, etc...
- Want to find games for your kids?
- Want to find a movie for the family?
- Want to buy holiday gift(s)?
Proactive IR in Mobile or Car

- Physical limitation of the mobile device limits a user’s ability/intention to search
  - Proactive information retrieval is needed
- Large amount of personal and contextual information available
  - Proactive information retrieval is likely to do well
Beyond Siri: Personal Assistant

- Open domain
- Don’t suddenly switching modes when can not handle a hard environment

Augmented reality through electric contact lens or glasses

Whisper in your ear
Major Challenges & Outline

- What to recommend
  - Will this user like the item (document, movie, products, restaurant, etc.)?
  - Use rich user interaction sequences, multi-model sensor data
  - Context: time and location
  - Consider the amount of time or information required
  - Open domain

- When to recommend: interruptibility; value of the information changes over time

- How to recommend: trust, explanation, educating a user, helping a user to choose
What to Recommend: Learn about the User

Forrest Gump:

"There is an awful lot you can tell about a person by their shoes ... where they're going, where they've been."
Find Your Soul Mate in Less than 10 Minutes

You can tell alot about people by their reactions and opinions about other people.
-8 Minute Dating
Our Approach: Developing System with Desirable Characteristics

What can a person do? (desirable characteristics)

- Use heuristics
- Ask good questions
- Use context and implicit feedback
- Social learning

Our solution for a computer

- Bayesian Prior
- Bayesian Active Learning
- Graphical Models
- Bayesian Hierarchical Models, Probabilistic Relational learning
What I Teach

- ISM260 Information Retrieval
- ISM58 System Analysis and Design (Winter)
- ISM245 Data Mining (Spring)
- Graduate seminar
Interested?

- Undergraduate research opportunities
  - Good at programming and math/statistics
  - Eligible for work study or National Science Foundation scholarship

- Try goodbuylist.com and give me your feedback

**GoodBuyList Contest!**

Recommend the best product, and win an iPad mini

*see rules for details*