Introduction to Search/Retrieval Technologies

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

Which Search Engine Do You Use?

What is Traditional Information Retrieval?
- A collection of unstructured 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
- Search
  - Human computer interaction
  - Computer networks
  - Security
- Database
  - Systems
- Applications
  - Bioinformatics
  - Business applications
  - Medical informatics
  - Digital libraries
  - Web search
  - Online marketing
  - Online advertising

Major Technologies: Go Beyond Traditional Search
- Access Information
  - Categorization
  - Filter
  - Search
- Generate Knowledge
  - Clustering
  - Information Retrieval
  - Question and Answer
  - Information Extraction
  - Text mining

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

- 3rd Century B.C. Library of Alexandria
  - 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 Fiedler

A Typical/Simple Retrieval Process

-Based on Bruce Croft and Ned Fiedler

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

A Typical Crawler Infrastructure
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 PageRank: 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 you will use to fight against the spammers.

Key Research Directions

- Natural Language Processing
- Personalization
- Proactive search

What is NLP?

- Arabic text: البعيد على الإنسان أن يكون أميراً ومستقلاً مع نفسه ومع أباه وجدته، وأن ينتسح... كأنه في عالم من الوحي وأن يترك على ما... 
- How can a computer make sense out of this string?
  - Morphology: What are the basic units of meaning (words)?
  - 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 Changxiang Zhai
You need to know new products related to yours
You want to know good restaurants in Santa Cruz
E.g., from the literature, web, or discussion groups,
You want good movies for your family

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
buy.com
CNET
Ask
ebay
Yahoo!

---

**What Machine Can do Now?**

The State of the Art

- Parsing: partial >90% (?)
- Speech act analysis: ???
- Pragmatic analysis: ????
- Anaphora resolution
- Word sense disambiguation
- Entity/relation extraction
- Anaphora resolution

From Chengxiang Zhai

---

**News/Messages/Discussions About Your Stock?**

---

**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**

Build Proactive Information Retrieval System with Desirable Characteristics

<table>
<thead>
<tr>
<th>What can a software with human intelligence do? (desirable characteristics)</th>
<th>Our solution for a computer</th>
<th>Unified Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use heuristics</td>
<td>Bayesian Prior</td>
<td></td>
</tr>
<tr>
<td>Ask good questions</td>
<td>Bayesian Active Learning</td>
<td>Bayesian Graphical Models</td>
</tr>
<tr>
<td>Use context and implicit feedback</td>
<td>Graphical Models</td>
<td></td>
</tr>
<tr>
<td>Learn from other users</td>
<td>Bayesian Hierarchical Models</td>
<td></td>
</tr>
</tbody>
</table>

A Special Browser With Feedback Interface

**Evidence Collected**

<table>
<thead>
<tr>
<th>Examples</th>
<th>Explicit user feedback</th>
<th>Relevancy, novelty, authority, readability, “user like”</th>
</tr>
</thead>
<tbody>
<tr>
<td>User actions</td>
<td>mouse usage, time on page, scroll bar usage, keyboard usage</td>
<td></td>
</tr>
<tr>
<td>Topic info.</td>
<td>topic familiarity</td>
<td></td>
</tr>
<tr>
<td>News source information</td>
<td>in link count</td>
<td></td>
</tr>
<tr>
<td>Content of each document</td>
<td>relevance score, readability score and document length</td>
<td></td>
</tr>
</tbody>
</table>
**What Machine Can Learn From the Data Collected**

User independent:

User specific:

---

**Other Research Projects in My Group**

- Adaptive Information filtering
  - news, blogs, movies, tv
- Personalized search
- Petascale Distributed search (with SSRC)
- Text mining from web, blog, call center data
  - For economy research (with Economist)
  - For product reviews

---

**Sponsors of Students in My Lab**

NEC

Nokia

Google

Los Alamos National Laboratory

---

**Highlight on Our Work**

**What I Teach**

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

---

**Interested?**

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