Web Information Retrieval

Textbook by
Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze
Notes Revised by X. Meng for SEU
May 2014

Acknowledgement

Contents of lectures, projects are extracted and organized from many sources, including those from Professor Manning's lecture notes from the textbook, notes and examples from others including Professor Bruce Croft of UMass, Professor Raymond Mooney of UT Austin, Professor David Yarowsky of Johns Hopkins University, Professor David Grossman of IIT, and Professor Brian Davidson of Lehigh.

Topics Covered By The Course

- Information retrieval models
- Retrieval evaluations
- Text properties
- Indexing and searching
- Web search engine architecture

Learning Outcomes

- After completing the course, students will be able to
  - Explain basic information retrieval models such as vector space model and probabilistic model
  - Evaluate the performance of information retrieval systems
  - Analyze information retrieval systems such as web search engine using the principles of IR
  - Design and implement a simple web search engine

Expected Background

- Data structures
- Programming
- Some network concepts (client, server, TCP/IP)
- The programming project in this course is a good culminating experience where one can put all computer science knowledge into the project

Search and Information Retrieval

- Search on the web is a daily activity for many people throughout the world
- Search and communication are most popular uses of the computer
- Applications involving search are everywhere
- The field of computer science that is most involved with R&D for search is *information retrieval (IR)*

Information Retrieval

- *"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."* (Salton, 1968)
- General definition that can be applied to many types of information and search applications
- Primary focus of IR since the 50s has been on *text* and *documents*
- In recent years the focus has been shifting towards multimedia (audio and video)

Web Information Retrieval

- While information retrieval (IR) generally applies to any kind of documents, web information retrieval works of the documents on the web
- Features specific to web documents
  - Mostly not structured
  - Huge quantity
  - Dynamic changes

8

What is a Document?

- Examples:
  - web pages, email, books, news stories, blog posts, scholarly papers, text messages, Word™, PowerPoint™, PDF, forum postings, patents, IM sessions, etc.
- Common properties
  - Significant text content
  - Some structure (e.g., title, author, date for papers; subject, sender, destination for email)
  - More recently: dynamic, program generated

Documents vs. Database Records

- Database records (or *tuples* in relational databases) are typically made up of well-defined fields (or *attributes*)
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics to queries in order to find matches
- Text is less structured, thus more difficult to work with

Search in Documents vs. Find Records

- Example bank database query
  - *Find records with balance > $50,000 in branches located in Amherst, MA.*
  - Matches easily found by comparison with field values of records
- Example search engine query
  - *Boston Marathon Bombing on April 16th, 2013*
  - Which word should be used for search?

Comparing Text

- Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval
- Exact matching of words is not enough
  - Many different ways to write the same thing in a "natural language" like English or Chinese
  - e.g., does a news story containing the text "*Boston Marathon Explosion*" match the query?
  - How about *"Marathon in London April 21st, 2013"*
  - Some stories will be better matches than others

Boundaries of Words

- In English words are separated by delimiters such as space
- In Chinese (and other languages), it is a bit more challenge to separate words (or characters)

和尚

The two characters can be treated as one word meaning 'monk' or as a sequence of two words meaning 'and' (和, 以及) and 'still, not yet (尚未).'

Chinese: No White Space

莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

Discussion

- What are some of the popular Chinese search engines?
- What are common ways of using a Chinese search engine?
- Other systems that act in a similar way to, or use extensively of, search engines? (e.g., QQ, Sina WeiBo?)

15

Dimensions of IR

- IR is more than just text, and more than just web search
  - although these are central
- People doing IR work with different media, different types of search applications, and different tasks

Other Media

- New applications increasingly involve new media
  - e.g., video, photos, music, speech
- Like text, content is difficult to describe and compare
  - text may be used to represent them (e.g., tags)
- IR approaches to search and evaluation are appropriate

Dimensions of IR

Content	Applications	Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned docs	Desktop search	Question answering
Audio	Forum search	
Music	P2P search	
	Literature search	

IR Tasks

- Ad-hoc search
  - Find relevant documents for an arbitrary text query
- Filtering
  - Identify relevant user profiles for a new document
- Classification
  - Identify relevant labels for documents
- Question answering
  - Give a specific answer to a question

Big Issues in IR

- Relevance
  - What is it?
  - Simple (and simplistic) definition: A relevant document contains the information that a person was looking for when they submitted a query to the search engine
  - Many factors influence a person's decision about what is relevant: e.g., task, context, novelty, style
  - *Topical relevance* (same topic) vs. *user relevance* (everything else)

Big Issues in IR

- Relevance
  - *Retrieval models* define a view of relevance
  - *Ranking algorithms* used in search engines are based on retrieval models
  - Most models describe statistical properties of text rather than linguistic
    - i.e., counting simple text features such as words instead of parsing and analyzing the sentences
    - Statistical approach to text processing started with Luhn in the 50s
    - Linguistic features can be part of a statistical model

Big Issues in IR

- Evaluation
  - Experimental procedures and measures for comparing system output with user expectations
    - Originated in Cranfield experiments in the 60s
  - IR evaluation methods now used in many fields
  - Typically use *test collection* of documents, queries, and relevance judgments
    - Most commonly used are TREC collections
  - *Recall* and *precision* are two examples of effectiveness measures

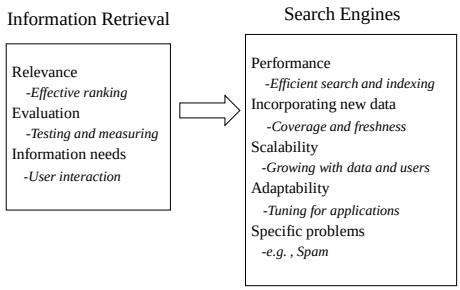
Big Issues in IR

- Users and Information Needs
  - Search evaluation is user-centered
  - Keyword queries are often poor descriptions of actual information needs
  - Interaction and context are important for understanding user intent
  - Query refinement techniques such as *query expansion*, *query suggestion*, *relevance feedback* improve ranking

IR and Search Engines

- A search engine is the practical application of information retrieval techniques to large scale text collections
- Web search engines are best-known examples, but many others
  - *Open source* search engines are important for research and development
    - e.g., Lucene, Lemur/Indri, *Galago*
- Big issues include main IR issues but also some others

IR and Search Engines



Search Engine Issues

- Performance
  - Measuring and improving the efficiency of search
    - e.g., reducing *response time*, increasing *query throughput*, increasing *indexing speed*
  - *Indexes* are data structures designed to improve search efficiency
    - designing and implementing them are major issues for search engines

Search Engine Issues

- Dynamic data
  - The "collection" for most real applications is constantly changing in terms of updates, additions, deletions
    - e.g., web pages
  - Acquiring or "crawling" the documents is a major task
    - Typical measures are *coverage* (how much has been indexed) and *freshness* (how recently was it indexed)
  - Updating the indexes while processing queries is also a design issue

Search Engine Issues

- Scalability
  - Making everything work with millions of users every day, and many terabytes of documents
  - Distributed processing is essential
- Adaptability
  - Changing and tuning search engine components such as ranking algorithm, indexing strategy, interface for different applications

Spam

- For Web search, spam in all its forms is one of the major issues
- Affects the efficiency of search engines and, more seriously, the effectiveness of the results
- Many types of spam
  - e.g., spamdexing or term spam, link spam, "optimization"
- New subfield called *adversarial IR*, since spammers are "adversaries" with different goals

Course Goals

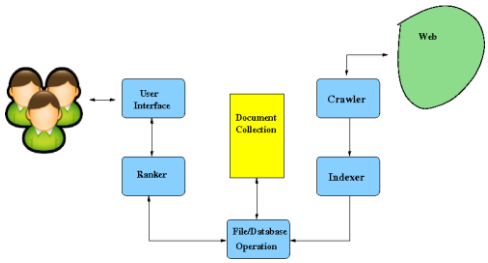
- To help you to understand search engines, evaluate and compare them, and implement a simple search engine
- Provide broad coverage of the important issues in information retrieval and search engines
  - includes underlying models and current research directions

Course Project Discussion

SEME: Search Engine Made Easier

31

Search Engine Architecture



32

Phased Approach

- Phase 1: Building a web server with your own home page(s) and user interaction
- Phase 2: Text processing and indexing
- Phase 3: Crawling the web
- Phase 4: Putting all together as a Boolean search engine (no ranking)
- Phase 5: Ranking search results (time permitting)

33

Some Web Statistics and History

34

Number of Web Servers

- According to the April 2014 Netcraft survey, there are 958,919,789 host names, 39 millions more than previous month
  - http://news.netcraft.com/archives/category/web-server-survey/
  - Of which 182,138,695 were active
  - Compared to the May 2013 statistics:
    - 672,837,096 host names
    - About 186 millions were active

35

Web Server Market Share

- According to the April 2014 Netcraft survey,
  - http://news.netcraft.com/archives/category/web-server-survey/
  - Of all the servers, Apache is the most popular one, 361,853,003 or 37.74%, followed by Microsoft servers, 316,843,695 or 33.04%

36

Statistics from Other Sources

- As of July 2012, there are over 900 million hosts on the Internet
  https://www.isc.org/solutions/survey/history
- After meeting someone they are interested in, 26% of US adults Google them (April 2013)
  - http://www.factbrowser.com/facts/11787/
- Google sites account for 67.1% of US searches, compared to 16.9% for Microsoft, and 11.8% for Yahoo! (March 2013)
  - http://www.factbrowser.com/facts/11410/

37

Related Areas

- Database Management
- Library and Information Science
- Artificial Intelligence
- Natural Language Processing
- Machine Learning

38

Database Management

- Focused on *structured* data stored in relational tables rather than free-form text.
- Focused on efficient processing of well-defined queries in a formal language (SQL).
- Clearer semantics for both data and queries.
- Recent move towards *semi-structured* data (XML) brings it closer to IR.

39

Library and Information Science

- Focused on the human user aspects of information retrieval (human-computer interaction, user interface, visualization).
- Concerned with effective categorization of human knowledge.
- Concerned with citation analysis and *bibliometrics* (structure of information).
- Recent work on *digital libraries* brings it closer to CS & IR.

40

Natural Language Processing

- Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse.
- Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on *meaning* rather than keywords.

41

Natural Language Processing: IR Directions

- Methods for determining the sense of an ambiguous word based on context (*word sense disambiguation*).
- Methods for identifying specific pieces of information in a document (*information extraction*).
- Methods for answering specific NL questions from document corpora.

42

Artificial Intelligence

- Focused on the representation of knowledge, reasoning, and intelligent action.
- Formalisms for representing knowledge and queries:
  - First-order Predicate Logic
  - Bayesian Networks
- Recent work on web ontologies and intelligent information agents brings it closer to IR.

43

Machine Learning

- Focused on the development of computational systems that improve their performance with experience.
- Automated classification of examples based on learning concepts from labeled training examples (*supervised learning*).
- Automated methods for clustering unlabeled examples into meaningful groups (*unsupervised learning*).

44

Machine Learning: IR Directions

- Matrix decomposition
  - Reduce higher dimension matrix to the ones that are "manageable," yet keep the semantics
- Community detection and interaction
  - Network science, social groups and their inter- or intra-influence
- Text Mining
  - Find patterns in text

45

Web Challenges for IR

- Distributed Data: Documents spread over millions of different web servers.
- Volatile Data: Many documents change or disappear rapidly (e.g., dead links).
- Large Volume: Billions of separate documents.
- Unstructured and Redundant Data: No uniform structure, HTML errors, up to 30% (near) duplicate documents.
- Quality of Data: No editorial control, false information, poor quality writing, typos, etc.
- Heterogeneous Data: Multiple media types (images, video), languages, character sets, etc.

46