What is Information Retrieval (IR)?

IR is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the WWW.

IR: The techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system.

Web Search

- Very similar to information retrieval
- Main differences:
  - Links between Web pages can be exploited
  - Collecting, storing, and updating documents is more difficult
  - Usually, the number of users is very large
  - Spam is a problem

Information Retrieval vs. Databases

<table>
<thead>
<tr>
<th>Information retrieval</th>
<th>Data retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieve all objects relevant to some information need</td>
<td>Retrieve all objects satisfying some clearly defined conditions</td>
</tr>
<tr>
<td>Find all documents about the topic “semantic web”?</td>
<td>SELECT id FROM document WHERE title LIKE ‘%semantic web%’</td>
</tr>
<tr>
<td>Result list</td>
<td>Well-defined result set</td>
</tr>
</tbody>
</table>

Why Should I Know about All This?

- “80% of business is conducted on unstructured information”
- “85% of all data stored is held in an unstructured format”
- “7 million Web pages are being added every day”
- “Unstructured data doubles every three months”
Why Should I Know About All This?

- Managing the information flood
- Have you ever tried to drink from a fire hydrant?

Organizational Issues

- Course overview
  - 13 lectures
  - Exercises are integrated into lectures
  - Wednesdays, 10:00–12:20 (including a 5-minute break)
- Final exam (for Bachelor’s and Master’s students)
  - Oral exam

Homework

- Homework exercises will be published every week
- However, there will be no grading
  - That is, homework is optional
- Discussion of solutions at the beginning of each lecture
- Sometimes, there will be practical exercises
  - Idea: Give you an impression how algorithms really work

Hints for exam preparation:

- Work together and discuss the lecture’s topics in small groups
- Do all homework exercises

Course Overview

1. Introduction and fundamental notions
2. Retrieval models: fuzzy, coordination level matching, vector space
3. Probabilistic retrieval models
4. Indexing
5. Latent Semantic Indexing
6. Language models, retrieval evaluation
7. Document clustering
8. Relevance feedback, classification
9. Support vector machines
10. Introduction to Web retrieval
11. Web crawling
12. Link analysis
13. Miscellaneous

Literature

http://www.informationretrieval.org


http://www.dcs.gla.ac.uk/Keith/Preface.html
Lecture 1: Introduction

1. A Brief History of Libraries, Information Retrieval, and Web Search
2. Fundamental Notions
3. IR Systems and Models
4. The Boolean Retrieval Model

Ancient Libraries

- Sumerian archives
  - Around 3000–2000 BC
  - About 25,000 clay tablets stored in temple rooms
  - Mostly inventories and records of commercial transactions
- The Great Library of Alexandria
  - Founded about 300 BC
  - Idea: A universal library holding copies of all the world’s books
  - At its height, the library held nearly 750,000 scrolls

Medieval Libraries

- Monastic libraries
  - Educated monks saved many ancient texts from getting lost by hand-copying
  - The Vatican Library was formally founded in 1475 but is in fact much older
- Gutenberg’s printing press
  - Around 1450, Johannes Gutenberg introduced movable type to Europe
  - The technique spread rapidly, copying books became much easier and less expensive

Modern Libraries

- German National Library
  - 25 million items
  - Located in Leipzig, Frankfurt (Main), and Berlin
- Library of Congress
  - 150 million items
  - 20 million new items since 2009
  - The world’s largest library (according to the Guinness Book)
  - Classification system: Library of Congress Classification

Library Catalogs

Items are cataloged by metadata:

- Author/Editor, ISBN, …
- Keyword, e.g. “information retrieval”
- Subject area, e.g. “information systems”
- Specialized classification systems, e.g. Library of Congress

MEDLINE and MeSH

- MEDLINE
  - Medical Literature Analysis and Retrieval System Online
  - A literature database of life sciences and biomedical information
  - Compiled by the U.S. National Library of Medicine
  - More than 18 million records from approximately 5,000 selected publications, ranging over more than 40 years
  - Manually indexed using a controlled vocabulary (leading to about 12 keywords assigned to each document)
  - Freely available on the Internet via PubMed: http://pubmed.gov
MEDLINE and MeSH (2)

- MeSH
  - Medical Subject Headings
  - MEDLINE’s controlled vocabulary
  - Around 25,000 subject headings (descriptors, keywords)
  - Arranged in a hierarchy
  - Also contains about 140,000 supplementary concept records, largely synonyms

Dublin Core Metadata (2)

- DCMI
  - Dublin Core Metadata Initiative
  - Standard for cross-domain information resource description
  - Defined in ISO Standard 15836
  - “Dublin” refers to Dublin, Ohio (location of a workshop in 1995)
  - 15 core metadata elements:
    - Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights

Full Text Search?

- Catalogue cards are document proxies
  - Often, they suffice to judge the relevance of a particular item for your information need
  - But:
    - A clever classification scheme is required:
      - Extensive enough to allow detailed classification
      - Simple enough to be easily understandable
    - Experts must catalogue each item individually
  - Problem: A lot of manual work!
  - Full text search: Every word is a keyword!

Pre-computer area: Concordances

- Alphabetical list of the principal words used in a book
- Only for works of special importance, such as the Bible
- First Bible concordance by Hugo de Saint Charo, with the help of 500 monks, around 1250

The Memex

- Vision of a hypertext-based PDA
  - Proposed by Vannevar Bush
    - Director of the Office of Scientific Research and Development (USA, 1941–1947)
  - Outlined in Bush’s famous essay “As We May Think” published in The Atlantic Monthly (1945)
  - “A device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility.”
  - “Selection by association, rather than by indexing.”
**The Memex (2)**

*Image of the Memex concept showing a desk with a notebook, a typewriter, and various other items.*

**Early Information Retrieval Systems**

- 1957: Hans-Peter Luhn (IBM) uses words as indexing units for documents
  - Measure similarity between documents by word overlap
- 1960s and 1970s: Gerard Salton and his students (Harvard, Cornell) create the SMART system
  - Vector space model
  - Relevance feedback

**IR Becomes a Research Discipline**

- ACM's SIGIR
  - Special Interest Group on Information Retrieval
  - Annual conferences, beginning in 1978
  - Gerald Salton award, first honoree: Gerald Salton (1983)
- TREC
  - Annual Text Retrieval Conference, beginning in 1992
  - Sponsored by the U.S. National Institute of Standards and Technology as well as the U.S. Department of Defense
  - Today: many different tracks, e.g., blogs, genomics, spam
  - Provides data sets and test problems

**A Brief History of Web Search**

- First Web search engines:
  - Archie: Query file names by regular expressions
  - Architext/Excite: Full text search, simple ranking (1993)
- Until 1998, web search meant information retrieval
- 1998: Google was founded
  - Exploits link structure using the PageRank algorithm

**Core Problems**

- How to store and update large document collections?
  - Small!
  - Scalable!
- How to do efficient retrieval?
  - Fast!
- How to do effective retrieval?
  - High result quality!

**Lecture 1: Introduction**

1. A Brief History of Libraries, Information Retrieval, and Web Search
2. Fundamental Notions
3. IR Systems and Models
4. The Boolean Retrieval Model
A document is a coherent passage of free text. "Coherent" means is about related topics. "Free" means natural, written language. Examples:
- Newspaper article
- Scientific article
- Dictionary entry
- Web page
- Email message

A document collection is a set of documents. Also known as corpus. Usually, all documents within a collection are similar with respect to some criterion. Examples:
- MEDLINE
- The articles covered by Google News
- The Web

An information need is the topic about which the user desires to know more. Refers to an individual, hidden cognitive state. Depends on what the users knows and doesn’t know. Ill-defined. Examples:
- What is the capital of Uganda?
- Is it really true that McDonald’s hamburgers contain worm meat?
- What is “cloud computing”?

A query is what the user conveys to the computer in an attempt to communicate the information need. Stated using a formal query language. Usually a list of search terms. But also: “Panda NEAR Jaguar BUT NOT animal”

A document is relevant with respect to some user’s information need if the user perceives it as containing information of value with respect to this information need. Usually assumed to be a binary concept, but could also be graded. Example:
- Information need: “What is relevance in IR?”
- Relevant document: Wikipedia’s entry “Relevance (information retrieval)”

Lecture 1: Introduction
1. A Brief History of Libraries, Information Retrieval, and Web Search
2. Fundamental Notions
3. IR Systems and Models
4. The Boolean Retrieval Model
Any IR system is based on an IR model. The model defines:

- ... a query language,
- ... an internal representation of queries,
- ... an internal representation of documents,
- ... a ranking function which associates a real number with each query-document pair.

Optional: A mechanism for relevance feedback.

The Bag of Words Model (2)

- Cons:
  - Word order gets lost
  - Very different documents could have similar representations
  - Document structure (e.g., headings) and metadata is ignored

- Pros:
  - Simple set-theoretic representation of documents
  - Efficient storage and retrieval of individual terms
  - IR models using the bag of words representation work well!

The Bag of Words Model (3)

- Any document can be represented by an incidence vector:

  Each document is represented by a bag (= multiset) of terms from a predefined vocabulary.

- Standard case:
  - Vocabulary = set of all the words occurring in the collection's documents
  - Each document is represented by the words it contains

That's one small step for a man, a giant leap for mankind

\[
\begin{bmatrix}
1 & 1 & 1 & 2 & 1 & 1 & 1 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

Information Retrieval and Web Search Engines — Wolf-Tilo Balke and Joachim Selke — Technische Universität Braunschweig

Lecture 1: Introduction

1. A Brief History of Libraries, Information Retrieval, and Web Search
2. Fundamental Notions
3. IR Systems and Models
4. The Boolean Retrieval Model
### Boolean Retrieval

- The simplest (and arguably oldest) IR model
- Documents = sets of words (index terms)
- Query language = Boolean expressions over index terms
- Binary ranking function, i.e. 0/1-valued
- Retrieval is based on membership in sets
  - “Find all documents indexed by the word ‘mankind’!”
  - “Find all documents indexed by the word ‘man’ or ‘mankind’!”

### Boolean Connectives

- **Boolean connectives:**
  - Conjunction
  - Disjunction
  - Negation

- **Truth table:**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### Example

- Document\textsubscript{1} = \{step, mankind\}
- Document\textsubscript{2} = \{step, China\}
- Query\textsubscript{1} = “step AND mankind”
  - Result set: Document\textsubscript{1}
- Query\textsubscript{2} = “step OR mankind”
  - Result set: Document\textsubscript{1}, Document\textsubscript{2}

### Query Processing

- Usually, documents are indexed by an inverted index
  - For each index term, the set of documents containing this term is pre-computed and stored on disk
  - This enables fast query processing

- Document collection:
  - Document\textsubscript{1} = \{step, mankind\}
  - Document\textsubscript{2} = \{step, China\}

- Inverted index:
  - step: Document\textsubscript{1}, Document\textsubscript{2}
  - mankind: Document\textsubscript{1}
  - China: Document\textsubscript{2}

### Boolean Queries in Practice

- Warning:
  - Exclusive use of negation will result in large result sets!
  - Query\textsubscript{3} = “NOT mankind”
- To match natural language better,
  - “BUT NOT” can be used instead of “AND NOT”
  - Query\textsubscript{4} = “step BUT NOT China”
- Use “OF” to search for subsets of a given size:
  - Query\textsubscript{5} = “2 of \{step, mankind, China\}”
  - Query\textsubscript{5} = (“step AND mankind” OR “step AND China”) OR (mankind AND China)

### Query Processing (2)

- Thanks to the inverted index, queries of the type “Show me all documents containing term X” can be answered quickly
- Also quick to compute: unions and intersections of sets
- Example:
  - result of “mankind AND step” = (result of “mankind”) \cap (result of “step”)
  - result of “mankind OR step” = (result of “mankind”) \cup (result of “step”)
- Idea: Convert all queries to conjunctive normal form or disjunctive normal form
Conjunctive normal form (CNF)
- A propositional formula is in CNF if it is a conjunction of clauses
  - A clause is a disjunction of literals
  - A literal is a variable or its negation
- Theorem: Any propositional formula can be converted into an equivalent formula that is in CNF

Disjunctive normal form (DNF)
- A propositional formula is in DNF if it is a disjunction of conjunctive clauses
  - A conjunctive clause is a conjunction of literals
- Theorem: Any propositional formula can be converted into an equivalent formula that is in DNF

Query 6 = "step AND ((China AND taikonaut) OR man)"

Conjunctive normal form (CNF):
Query6 \equiv "step AND (China OR man) AND (taikonaut OR man)"

Disjunctive normal form (DNF):
Query6 \equiv "(step AND China AND taikonaut) OR (step AND man)"

Simple query paradigm, easy to understand
- If all document representations are mutually distinct, any possible subset of documents can be retrieved by a suitable query
  \Rightarrow cut out the set of relevant documents
- But: This advantage is rather theoretical, since the "right" query usually is unknown

Pros
- Westlaw
  - Online legal research service for US law
  - Includes more than 23,000 databases of case law, state and federal statutes, administrative codes, law journals, newspapers …
  - Indexed by the West Key Number System, a master classification system of U.S. law
  - Until recently, Boolean search has been the default method
Example 1:

- **Information need:** Information on the legal theories involved in preventing the disclosure of trade secrets by employees formerly employed by a competing company
- **Query:** "trade secret" /s disclos! /s prevent /s employe!

Example 2:

- **Information need:** Requirements for disabled people to be able to access a workplace
- **Query:** disab! /p access! /s (work-site work-place) (employment /3 place)

In 2005, Boolean search was the default in Westlaw

Submitted queries average about ten words in length

Professionals often prefer Boolean search to other methods because of greater control and transparency

But: In 1994, experiments on a Westlaw subcollection found that free text queries produced better results for queries prepared by Westlaw’s own librarians

More retrieval models

- Fuzzy retrieval model
- Coordination level matching
- Vector space model