CS-7961: Topics in Information Retrieval Seminar

Prof. Ellen Riloff

Information Retrieval

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Examples:
- web search
- reference librarians
- patent applications
- legal case retrieval

Information Retrieval (IR)

- The user has an ***information need***.
- The user provides a ***query*** that describes the information need.
- The IR system retrieves a set of documents from a ***corpus*** (document collection) that are believed to be ***relevant***.
- The documents are often ranked based on the likelihood that they are relevant.

IR Tasks

- The most familiar task is ***ad-hoc retrieval***:
  user provides a query expressing an information need and system returns relevant documents.
- *Text Classification/Categorization*: assign topic labels to documents (presumption of ongoing information need).
- *Text Filtering/Routing*: flag documents according to a profile either for routing (e.g., to an appropriate person) or for removal (e.g., spam, porn filtering).
- *Clustering*: organize a document collection by grouping similar or related documents.
  - *Information Visualization* is a growing need to visually represent the contents of extremely large document collections.

Types of Documents

- Unstructured: natural language text
  - There is linguistic structure, but little (if any) surface-level document structure.
- Semi-structured: some natural language text, but also some surface-level document structure.
  - Examples: resumes, seminar announcements
- Structured: data whose meaning is derived from the way it is organized
  - Databases are a common form of structured data.

Semi-Structured Text Examples

Seminar Announcement →

Laura Petitte
Department of Psychology
McGill University

Thursday, May 4, 1995
12:00 pm
Baker Hall 355

Name: Dr. Jeffrey D. Hermes
Affiliation: Department of AutoImmune Diseases Research & Biophysical Chemistry, Merck Research Laboratories
Title: "MHC Class II: A Target for Specific Immunomodulation of the Immune Response"
Host/e-mail: Robert Murphy, murph@a.crf.cmu.edu
Date: Wednesday, May 3, 1995
Time: 3:30 p.m.
Place: Mellon Institute Conference Room

Boolean Keyword Systems

- The user provides a list of keywords that are likely to appear in relevant documents.
  - Example: to find documents about conspiracy theories involving the assassination of JFK, the user might list: *JFK, conspiracy, assassination*
- By default, most systems use a Boolean ***and*** operator, but advanced search options usually support additional Boolean operators.
  - Example: (AND (OR(JFK,Kennedy), conspiracy, (OR(assassination,murder,shooting)))

Major Issues in IR

- ***Polysemy****:* many words have multiple meanings.
- ***Synonymy****:* many words can mean the same thing.
- ***Size/Speed****:* IR systems must process huge volumes of text, with instantaneous response time.
- ***Broad Coverage***: IR systems must be able to handle queries about any topic whatsoever.

Basic Evaluation Measures

- ***Precision*** = percentage of returned documents that are truly relevant.
  - Intuition: hit rate. How often is the system correct?
- ***Recall*** = percentage of all relevant documents that the system finds.
  - Intuition: coverage. How much of the desired material is found?

Inverted Index

- Most IR systems use an ***inverted index (inverted file)*** to represent the documents in the collection.
- Each document is ***tokenized*** to identify indivudal ***terms*** (normalized tokens).
- A dictionary is created from the terms, and each term is linked to a list of documents that contain the term (***postings***).

Inverted Index Example

Assassination	→	d1, d5, d21, d73, d304, d511…
Conspiracy	→	d3, d4, d7, d54, d73, d288…
JFK	→	d2, d21, d50, d73, d183, d288…
Kennedy	→	d2, d5, d66, d89, d214, d288…

The inverted index may also contain:
• frequency count of each term
• positional information

Stop Words

- Most IR systems use a ***stop list (stop words)***, which typically consist of closed class words that do not contain much semantic information.
- Stop words are not included in the inverted index, which dramatically reduces its size.

Typical stop words:
Articles: a, an, the
Prepositions: of, to, from, by, with, for, at, in…
Modals: would, could, should, can, will, must…
etc.

Disadvantages of Stop Words

- Common strings can be used in uncommon ways.
  - Example: *"the"* can be a Vietnamese name
- Stop words can be crucial parts of a lexicalized phrase, title, or quote.
  - Example*: "to be or not to be"*
- Some stopwords, such as prepositions, can provide important information about relationships.
- Disk space is much cheaper than it used to be, so saving space may not be as important as it once was.

Stemming

- Many IR systems use ***stemming*** to match query terms with morphological variants in the documents.
  - Example:
    - assassinate
    - assassinated
    - assassinates
    - assassinating
    - assassination
    - assassinations

Problems observed with the Porter Stemmer

Incorrect Conflation	
organization	organ
doing	doe
generalization	generic
numerical	numerous
policy	police
university	universe
easy	easily
addition	additive
negligible	negligent
execute	executive

Errors of Omission	
European	Europe
analysis	analyze
matrices	matrix
noise	noisy
sparse	sparsity
explain	explanation
resolve	resolution
triangle	triangular
urgency	urgent
cylinder	cylindrical

IR is not just web search!

- There are some very important real-world challenges! For example:
  - —Legal Search. Some real Westlaw information needs:

    Information on the legal theories involved in preventing the disclosure of trade secrets by employees formerly employed by a competing company.

    Cases about the host's responsibility for drunk guests.
  - —Question Answering. NLP meets IR: most people really want computers to be able to return a specific answer to a question, not a set of documents.
  - —Current IR systems do reasonably well with precision (for simple queries), but recall is still a major problem!