Advanced information retrieval

Computer engineering department

Spring 2005
Subjects

1. Introduction
2. Models
3. Retrieval evaluation
4. Query languages
5. Query reformulation
6. Text properties
7. Text languages
8. Text processing
9. Information retrieval from the Web
Sources

5. Various Web sites.
6. Original material.
Motivation

- IR: representation, storage, organization of, and access to information items
- Focus is on the *user information need*
- *Information item:* Usually text, but possibly also image, audio, video, etc.
- Presently, most retrieval of non-text items is based on searching their textual descriptions.
- Text items are often referred to as *documents*, and may be of different scope (book, article, paragraph, etc.).
- User information need:
  - Find all docs containing information on college tennis teams which: (1) are maintained by a USA university and (2) participate in the NCAA tournament.
- Emphasis is on the retrieval of information (not data)
Motivation

- Data retrieval
  - which docs contain a set of keywords?
  - Well defined semantics
  - a single erroneous object implies failure!

- Information retrieval
  - information about a subject or topic
  - semantics is frequently loose
  - small errors are tolerated

- IR system:
  - interpret contents of information items
  - generate a *ranking* which reflects relevance
  - *notion of relevance* is most important
**Motivation**

- IR at the center of the stage
  - IR in the last 20 years:
    - classification and categorization
    - systems and languages
    - user interfaces and visualization
  - Still, area was seen as of narrow interest
  - Advent of the Web changed this perception once and for all
    - universal repository of knowledge
    - free (low cost) universal access
    - no central editorial board
    - many problems though: IR seen as key to finding the solutions!
Objectives

- Overall objective:
  - Minimize search overhead
- Measurement of success:
  - Precision and recall
- Facilitate the overall objective:
  - Good search tools
  - Helpful presentation of results
Minimize search overhead

Minimize overhead of a user who is locating needed information.

- **Overhead**: Time spent in all steps leading to the reading of items containing the needed information (query generation, query execution, scanning results, reading non-relevant items, etc.).

- **Needed information**: Either
  - Sufficient information in the system to complete a task.
  - All information in the system relevant to the user needs.
  - Example –shopping:
    - Looking for an item to purchase.
    - Looking for an item to purchase at minimal cost.
  - Example –researching:
    - Looking for a bibliographic citation that explains a particular term.
    - Building a comprehensive bibliography on a particular subject.
Measurement of success

- Two dual measures:
  - **Precision**: Proportion of items retrieved that are relevant.
    - $Precision = \frac{\text{relevant retrieved}}{\text{total retrieved}}$
    - $= \frac{|\text{Answer} \cap \text{Relevant}|}{|\text{Answer}|}$
  - **Recall**: Proportion of relevant items that are retrieved.
    - $Recall = \frac{\text{relevant retrieved}}{\text{relevant exist}}$
    - $= \frac{|\text{Answer} \cap \text{Relevant}|}{|\text{Relevant}|}$
  - Most popular measures, but others exist.
Measurement of success (cont.)
Support user search

Support user search, providing tools to overcome obstacles such as:

- Ambiguities inherent in languages.
  - Homographs: Words with identical spelling but with multiple meanings.
  - Example: *Chinon*—Japanese electronics, French chateau.

- Limits to user's ability to express needs.
  - Lack of system experience or aptitude.

- Lack of expertise in the area being searched.
  - Initially only vague concept of information sought.
  - Differences between user's vocabulary and authors' vocabulary: different words with similar meanings.
Presentation of results

Present search results in format that helps user determine relevant items:

- Arbitrary (physical) order
- Relevance order
- Clustered (e.g., conceptual similarity)
- Graphical (visual) representation
Basic Concepts

The User Task

- Retrieval
  - information or data
  - purposeful

- Browsing
  - glancing around
  - F1; cars, Le Mans, France, tourism
Querying(retrieval) vs. Browsing

Two complementary forms of information or data retrieval:

- Querying:
  - Information need (retrieval goal) is focused and crystallized.
  - Contents of repository are well-known.
  - Often, user is sophisticated.
Querying (retrieval) vs. Browsing (cont.)

- Browsing:
  - Information need (retrieval goal) is vague and imprecise.
  - Contents of repository are not well-known.
  - Often, user is naive.

- Querying and browsing are often interleaved (in the same session).
  - Example: present a query to a search engine, browse in the results, restate the original query, etc.
Pulling vs. Pushing information

- Querying and browsing are both initiated by users (information is “pulled” from the sources).
- Alternatively, information may be “pushed” to users.
  - Dynamically compare newly received items against standing statements of interests of users (profiles) and deliver matching items to user mail files.
  - Asynchronous (background) process.
  - Profile defines all areas of interest (whereas an individual query focuses on specific question).
  - Each item compared against many profiles (whereas each query is compared against many items).
Basic Concepts

- Logical view of the documents

- Document representation viewed as a continuum: logical view of docs might shift