Contents

1 Introduction to Information Retrieval 1
   1.1 What Kind of Data does Information Retrieval Deal With? ........................ 1
   1.2 Why Do We Want to Study Information Retrieval? ................................... 1
   1.3 Why is Information Retrieval a Difficult Problem? ................................... 1
   1.4 Document Retrieval Model ................................................................. 2
   1.5 Objectives ......................................................................................... 2
   1.6 Course Outline .................................................................................. 3

1 Introduction to Information Retrieval

1.1 What Kind of Data does Information Retrieval Deal With?

- Non-formatted data (as opposed to relational database)
- Textual data: papers, technical reports, newspaper articles
- Non-textual data: images, graphics, video

Although most people feel rather distant from information retrieval systems, they actually have used these systems without knowing. Examples of information retrieval systems are: news systems, library systems, world-wide web, electronic encyclopedia, AnswerBook, and the UNIX man -k and grep commands. In the last case, the command grep keyword filenames constitutes a complete information retrieval system in that grep is the retrieval system, keyword is the query and filenames corresponds to the database. Try to identify the retrieval system, query, and database for the UNIX command: man -k xterm.

1.2 Why Do We Want to Study Information Retrieval?

- Most information available is in textual form and has no predefined format (e.g., emails and newsgroup articles).
- Integration of text retrieval capability in most relational database systems. SQL already supports limited search capability such as search based on regular expressions: Name like '%Lee%'.
- Increasing number of online documentation systems (no more hardcopy!).
- Of course, the blooming of World Wide Web.

1.3 Why is Information Retrieval a Difficult Problem?

- Huge amount of data (e.g., WWW) dictates efficiency, effectiveness and user-friendliness.
  A couple of decades ago, a few hundred Mbytes are considered large; today, a single database can easily grow to 10-50 Gbytes. Considering a distributed environment such as the WWW, the amount of data accessible is practically uncountable. According to some WWW search engines:
- 23,550 Web servers, 4 million URLs, 30 Gbytes (Lycos, April 1995).
- 30 million URLs, 40-Gbyte index (Alta Vista, May 1996), no mention of the total size of the pages.

- Unstructured data: difficult to capture semantics in documents. Compare:
  “select * from Employee where Salary > 100,000”
  “retrieve all news items about corporate takeover”

- Documents have unrestricted domains. For instance, it is hard to predefine or pre-categorize the subject domains of the documents.

- Diversified user base: expert to casual users.

- Intention of information and user query is hard to capture. Compare a README file and a user manual (And a summary versus an indepth report).

- Distributed and interlinked (e.g., Hypertext and WWW).
  Where to start a search? Unlike in a centralize database, you have only one (or a few) database(s) to search.
  How are the information related?

- Efficiency vs. effectiveness
  With a limited amount of resources, one can only improve efficiency and effectiveness to a certain degree. Moreover, improving efficiency often means degrading effectiveness, and vice versa.

1.4 Document Retrieval Model

![Diagram of Document Retrieval Model]

- We will cover in this course all of the boxes in this diagram.

1.5 Objectives

- Problems within the IR domain
- Theory and practices
- Hand-on experiences
- Future trend
- Advanced applications
1.6 Course Outline

- Introduction and course overview.
- Differences and similarity between database and information retrieval.
- Basic concepts of relational and object-oriented database; introduction to Illustra, a commercial object-relational DBMS.
- Information retrieval models, measure and evaluation of retrieval effectiveness.
- Boolean retrieval model.
- Document ranking: vector space model and probabilistic model.
- Relevance feedback techniques.
- Document clustering.
- Indexing techniques: inverted files and signature files.
- Text encoding methods.
- Network-based information retrieval systems.
- Advanced information retrieval techniques: fuzzy and knowledge-based information retrieval.
- Advanced applications: wireless information broadcast, digital library.