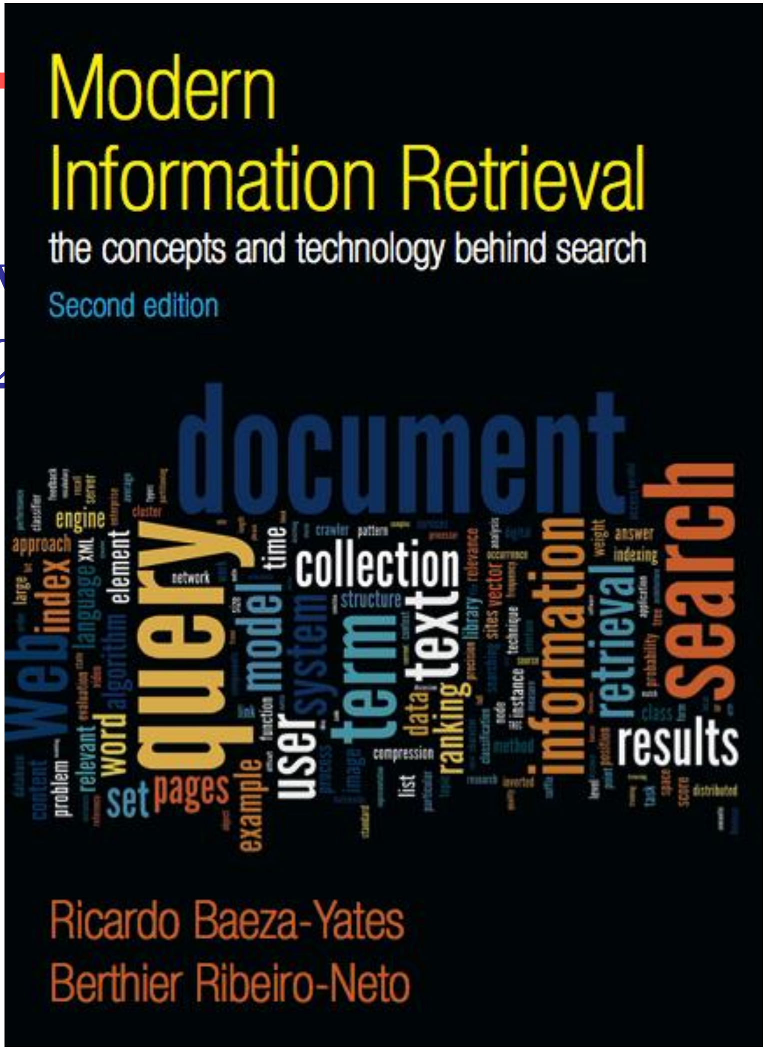
Information Storage and Retrieval

- Book
  - Modern information Retriev
    technology behind search (2
  - ISBN 978-0-321-41691
- Point
  - Assignment 30 %
  - Quiz 30%
  - Final 40%



Chapter 1
Introduction to IR

Motivation

- IR: representation, storage and access to information items
- Focus is on the *user information need*
- Emphasis is on the retrieval of information (not data)

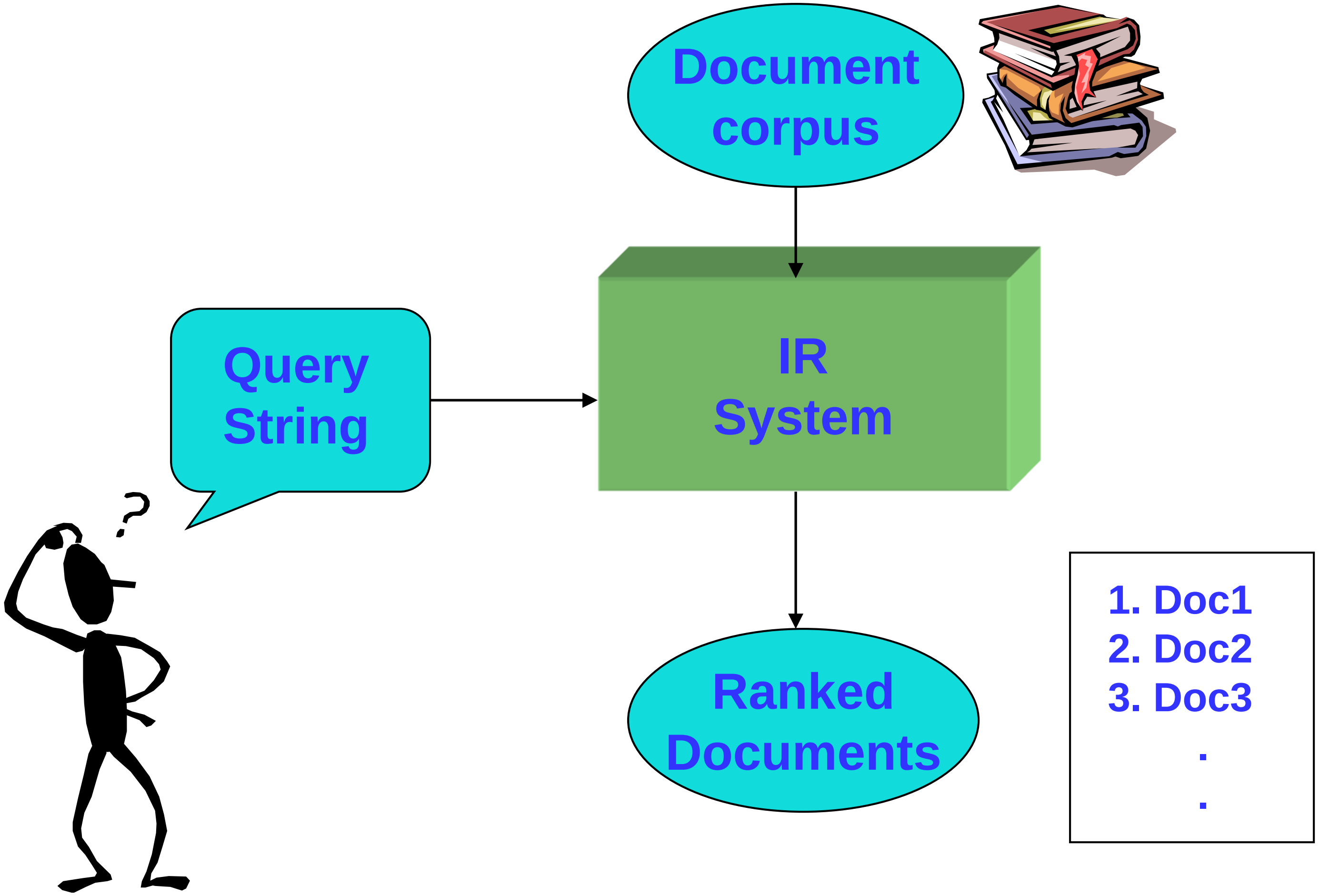
Comparing IR to databases

	Database	IR
Data	Structured	Unstructured
Fields	Clear semantics (SSN,age)	No fields (other than text)
Queries	Defined (relational algebra,SQL)	Free text(“natural language”),Boolean
Recoverability	Critical (Concurrency control,recovery, atomic operations)	Downplayed,though still an issue
Matching	Exact (results are always correct)	Imprecise (need to measure effectiveness)

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

IR System

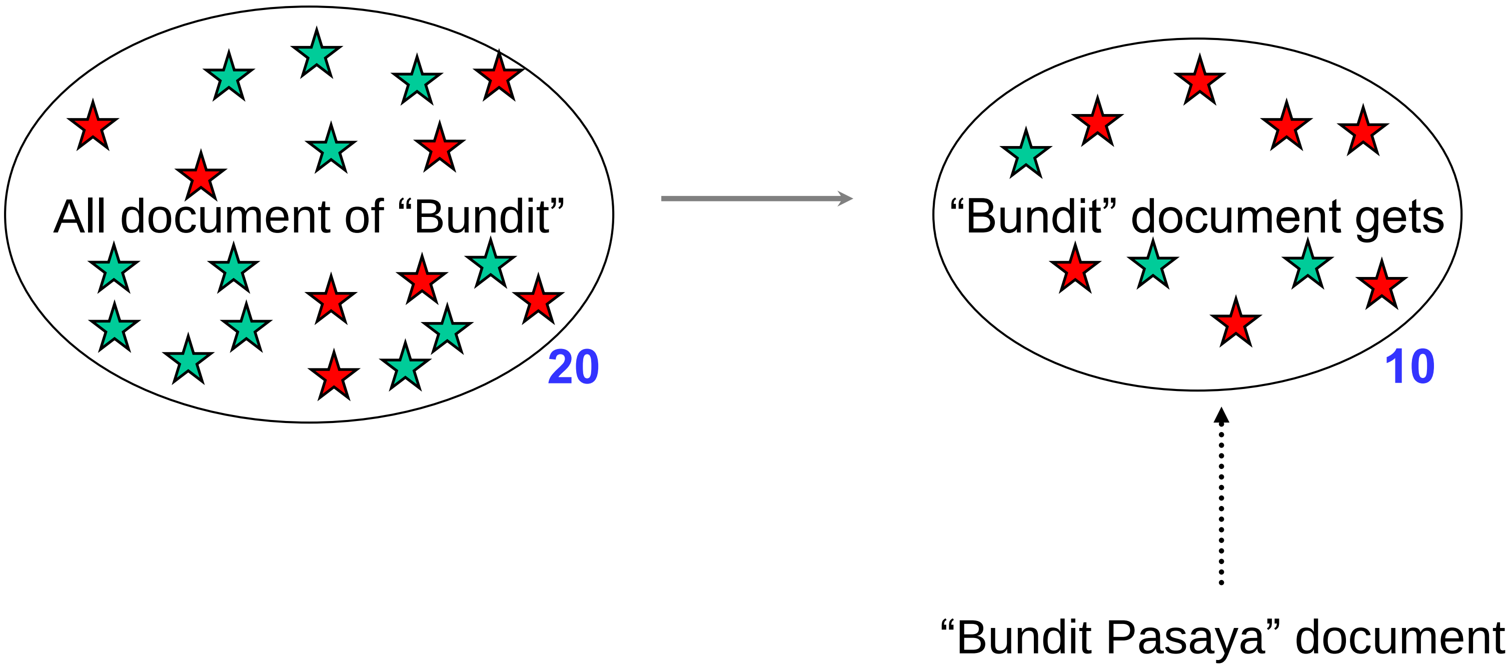


1. Doc1
2. Doc2
3. Doc3

.

.

Relevance Example



Relevance

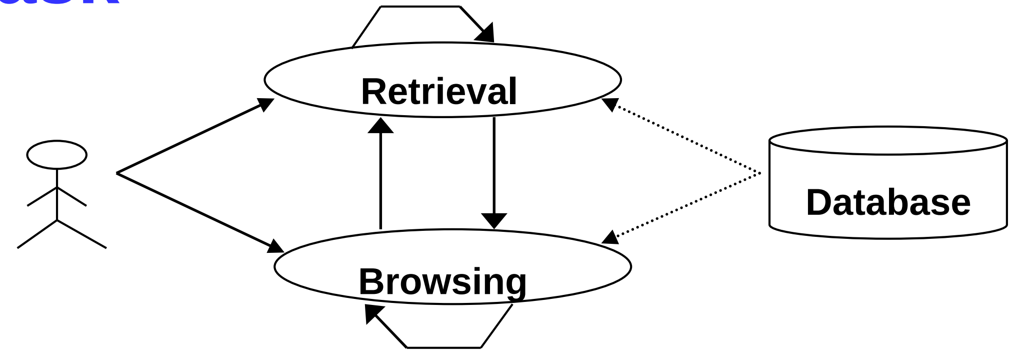
- Relevance is a subjective judgment and may include:
  - Being on the proper subject.
  - Being timely (recent information).
  - Being authoritative (from a trusted source).
  - Satisfying the goals of the user and his/her intended use of the information (***information need***).

Problems with Keywords

- May not retrieve relevant documents that include synonymous terms.
  - “restaurant” vs. “café”
  - “PRC” vs. “China”
- May retrieve irrelevant documents that include ambiguous terms.
  - “bat” (baseball vs. mammal)
  - “Apple” (company vs. fruit)
  - “bit” (unit of data vs. act of eating)

Basic Concepts

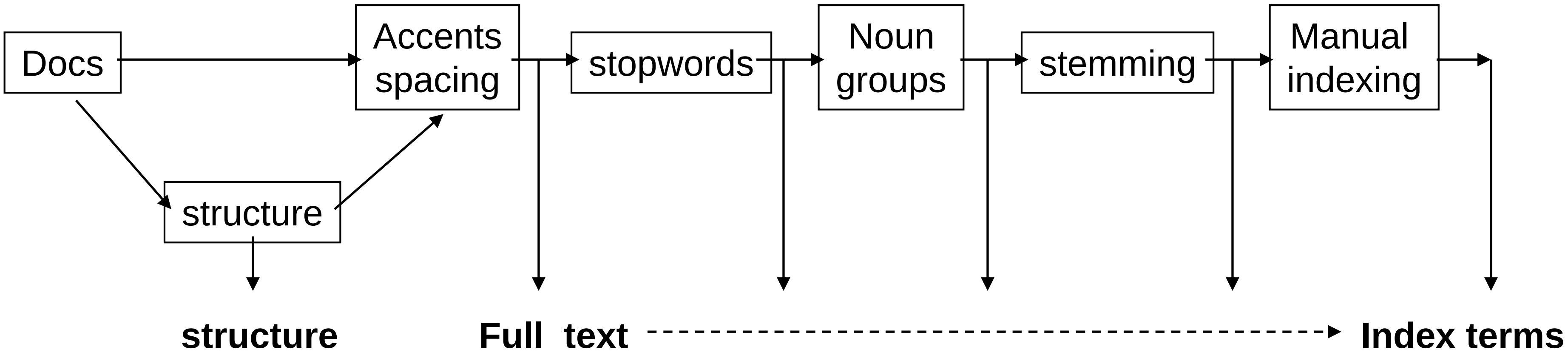
The User Task



- ◆ Retrieval
  - information or data
  - purposeful
- ◆ Browsing
  - glancing around
  - main objectives are not clearly defined in the beginnig
  - purpose might change during the interaction with system

Basic Concepts

- Logical view of the documents



IR Concepts

- Computer Center View
- Human Center View

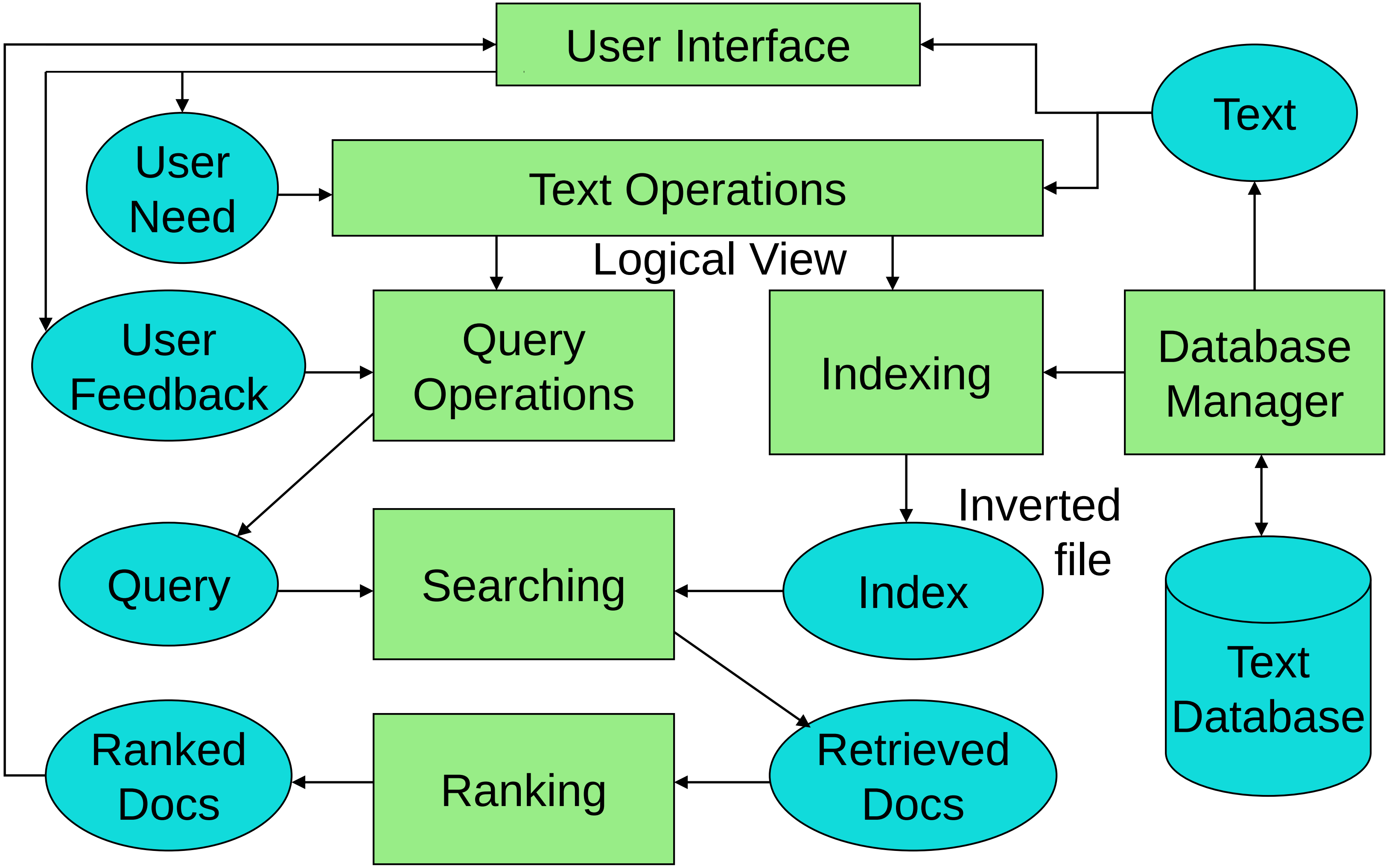
IR Questions

1. Translating user need
2. Using indices
3. Ranking

Recent IR History

- 2000’s continued:
  - Multimedia IR
    - Image
    - Video
    - Audio and music

IR System Architecture



IR System Components

- **Text Operations** forms index words (tokens).
  - Stopword removal
  - Stemming
- **Indexing** constructs an *inverted index* of word to document pointers.
- **Searching** retrieves documents that contain a given query token from the inverted index.
- **Ranking** scores all retrieved documents according to a relevance metric.

IR System Components (continued)

- **User Interface** manages interaction with the user:
  - Query input and document output.
  - Relevance feedback.
  - Visualization of results.
- **Query Operations** transform the query to improve retrieval:
  - Query expansion using a thesaurus.
  - Query transformation using relevance feedback.

Related Areas

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