1. Introduction to IR

Outline

1. Basic Concepts
    1. What is IR?
    2. Why IR?
    3. IR Task
    4. Information vs Data Retrieval
    5. Logical view of the document
2. The retrieval process
3. Taxonomy of IR
4. Classic IR and Alternative models

1. Basic Concepts

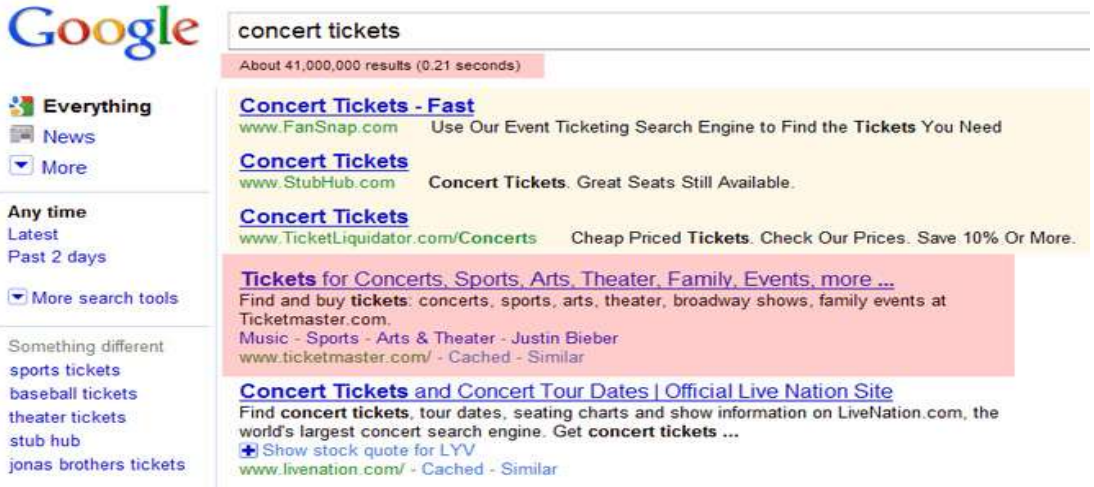
1.1 What is IR?

- 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).

- These days we frequently think first of web search, but there are many other cases:
  - E-mail search
  - Searching on your laptop
  - Corporate knowledge bases
  - Legal information retrieval

IR

1. Concerned firstly with retrieving *relevant* documents to a query.
2. Concerned secondly with retrieving from large sets of documents *efficiently*.
3. Ranking of search results



IR

An information retrieval model is a quadruple [D, Q, F, R(qi, dj)], where

- D is a set composed of logical views (or representations) for the documents in the collection
- Q is a set composed of logical views (or representations) for the user information needs. Such representations are called queries.
- F is a framework for modeling document representations, queries, and their relationships.
- R(qi, dj) is a ranking function which associates a real number with a query qi in Q and a document representation dj in D. Such ranking defines an ordering among the documents with regard to the query qi.

Types of Search Engine

- General: Google, Bing, Yahoo
- Business: Business.com, Thomasnet, GlobalSpec
- Education: Google Scholar, CiteULike
- Medical: Healthline, WebMD
- Similarly, law, jobs, legal, news, real estate etc.

1.2 Why IR?

- High velocity and massive amount of data (uploaded and deactivated)
- Types of websites
  - Social media websites, Social networking websites, Community based question answering, Forums, Blogs, Video sharing portals, News media, Image sharing websites, Online shopping websites, Bookmarking websites, Domain selling websites and many more…
- Problem with the data itself:
  - Distributed data
  - High percentage of volatile data
    - it is estimated that 40% of the Web changes every month.
  - Unstructured and redundant data
    - No conceptual model, no organization, no constraints.
    - By some estimates, about 30% of the Web is redundant.
  - Quality of data
    - Data can be false, invalid, outdated, poorly written or with many errors.
  - Heterogeneous data
    - Multiple media types, multiple formats, languages and alphabets.
- User and his interaction with the retrieval system:
  - How to specify a query
  - How to interpret the answer provided by the system.

1.3 IR Task

- **Input:**
  - A corpus of textual natural-language documents.
  - A user query in the form of a textual string.
- **Output:**
  - A ranked set of documents that are relevant to the query

Intelligent IR

- **Meaning of the words** used
- **Order of words** in the query
- **Direct or indirect feedback**
- **Authority** of the source

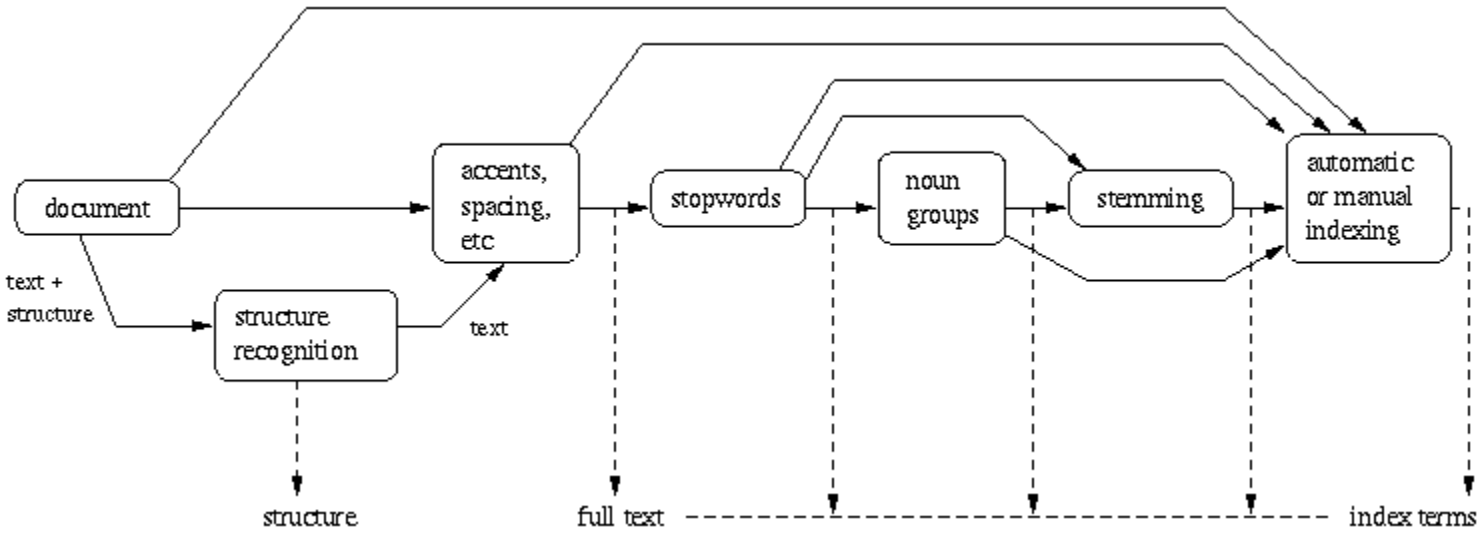
1.4 Data vs Information Retrieval

	Databases	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)	Downplayed, though still an issue
Matching	Exact (results are always “correct”)	Imprecise (need to measure effectiveness)

Data vs Information Retrieval

- **Data retrieval**
  - Which documents contain a set of keywords?
  - Well defined structure and semantics
  - A single erroneous object implies failure
  - Provide solution to the user of a database system
- **Information retrieval**
  - Information about a subject or topic
  - Semantics is frequently loose
  - Small errors are tolerated
  - Deals with natural language text

1.5 Logical View of the Document



from full text to a set of index terms

Logical View of the Document

- Documents in a collection are frequently represented through a set of index terms or keywords.
- Such keywords might be extracted directly from the text of the document or might be specified by a human subject (as frequently done in the information sciences arena).
- Modern computers are making it possible to represent a document by its full set of words.
- With very large collections, however, even modern computers might have to reduce the set of representative keywords.

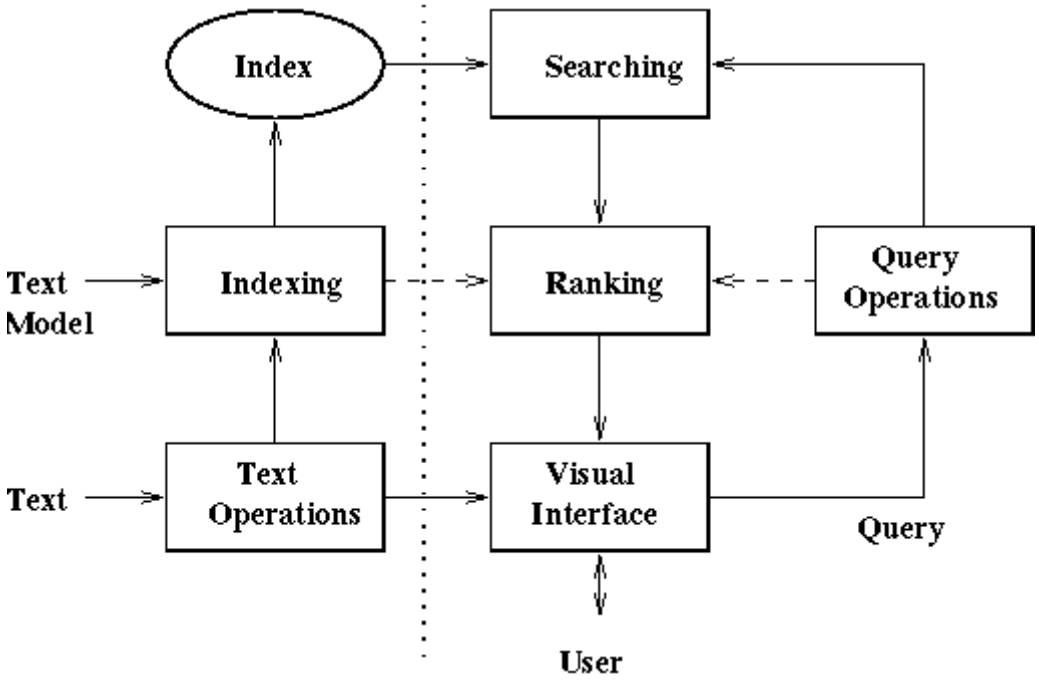
Logical View of the Document

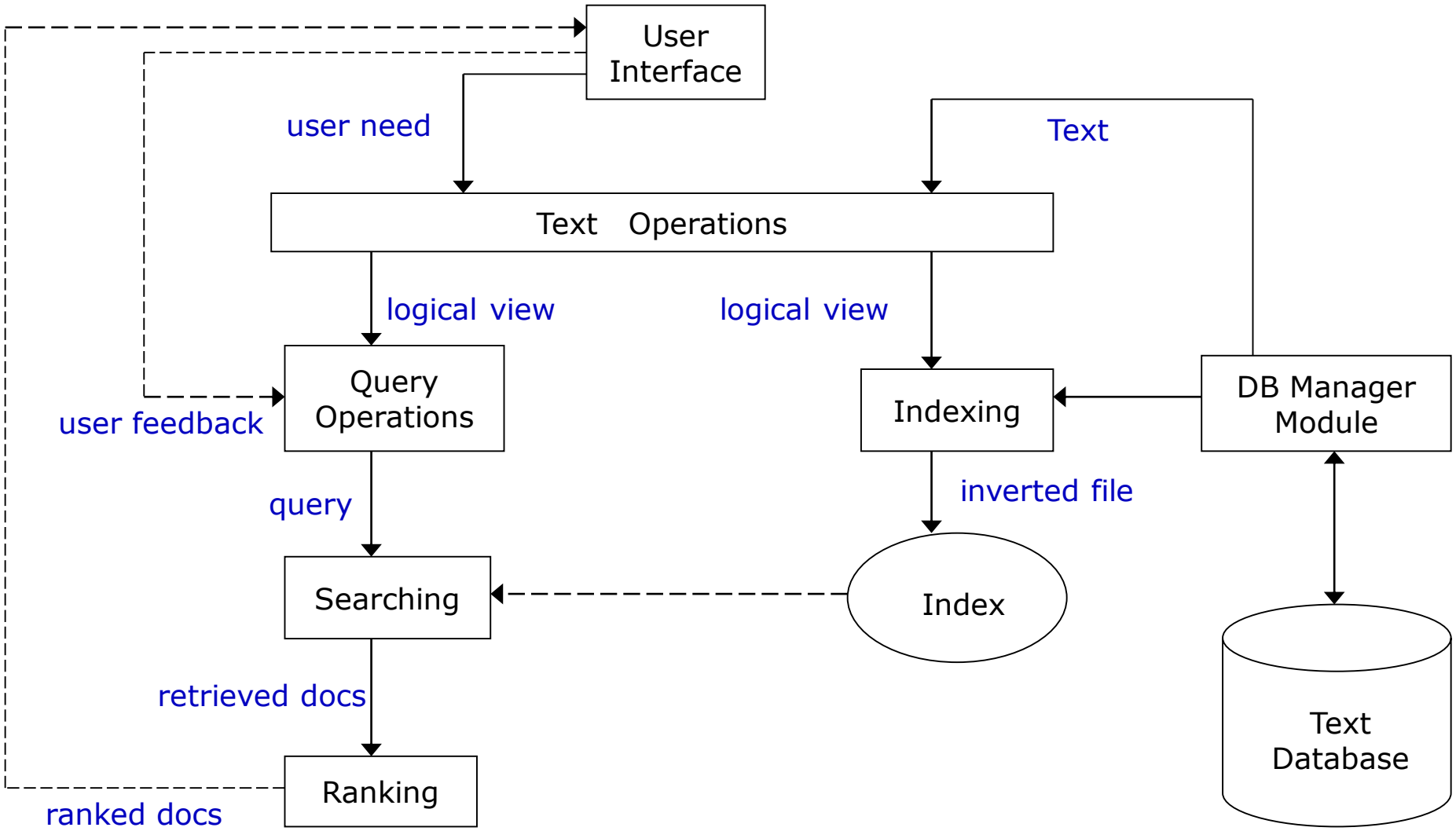
- This can be accomplished through
  - the elimination of **stopwords**
    - i.e. of, in, about, with, I, although, …
    - Stop-list: contain stop-words, not to be used as index
    - Prepositions, Articles, Pronouns
    - Some adverbs and adjectives, Some frequent words (e.g. document)
    - The removal of stop-words usually improves IR effectiveness
    - A few “standard” stop-lists are commonly used
  - the use of **stemming**
    - which reduces distinct words to their common grammatical root)
    - Different word forms may bear similar meaning: create a “standard” representation for them
    - **Computer, compute, computes, computing , computed , computation:= comput**
  - the identification of **noun groups**
    - which eliminates adjectives, adverbs, and verbs
- Further, compression might be employed. These operations are called text operations (or transformations).

Logical View of the Document

- The full text is clearly the most complete logical view of a document but its usage usually implies higher computational costs.
- A small set of categories (generated by a human specialist) provides the most concise logical view of a document but its usage might lead to retrieval of poor quality.
- Several intermediate logical views (of a document) might be adopted by an information retrieval system
- Besides adopting any of the intermediate representations, the retrieval system might also recognize the internal structure normally present in a document (e.g., chapters, sections, subsections, etc.).

1.2 The Retrieval Process




User Interface
user need
Text
Text Operations
logical view
logical view
Query Operations
user feedback
Indexing
DB Manager Module
query
inverted file
Searching
Index
retrieved docs
Text Database
Ranking
ranked docs

The Retrieval Process

- Before the RP can be initiated, it is necessary to define the text DB
- This is done the DB manager, which specifies the following,
  - The documents to be used
  - The operations to be performed on the text
  - The text model, i.e. the text structure and what elements can be retrieval
- Text operations transform the original documents and generate a logical view of them
- The database manager builds an index of the text i.e. “inverted file”
- Query operations used to generate actual “query” based on the user’s needs to retrieve the relevant document
- Before been sent to the user, the retrieved documents are ranked according to a likelihood of relevance

IR System Components

- Text Operations forms index words (tokens)
  - Stop-word 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
- 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
  - Query transformation using relevance feedback

1.3 Taxonomy of IR