Introduction to Information Retrieval

Günther Specht
Eva Zangerle

Summer Term 2019

Information Retrieval Definition

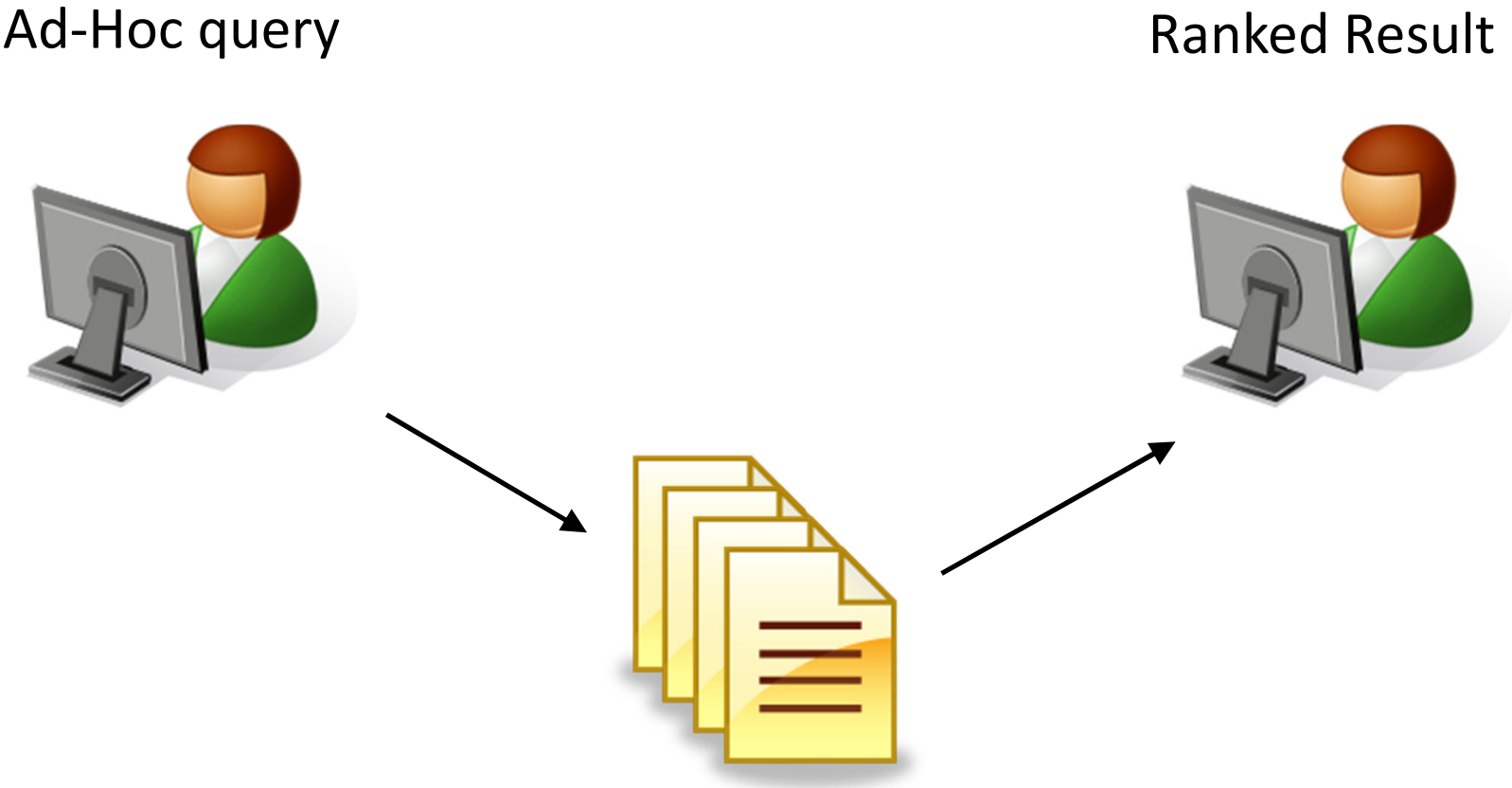
- Information retrieval (IR) deals with the *representation*, *storage*, *organization of*, and *access to information items*.
  (*Baeza-Yates, Ribeiro-Nieto, 1999)*

- Information retrieval (IR) is devoted to finding *relevant* documents, not finding simple match to patterns.
  (*Grossman - Frieder, 2004)*

- Information retrieval (IR) is finding material (usually documents) of an *unstructured nature* (usually text) that satisfy an *information need* within *large collections* (usually stored on computers).
  (M*anning et al., 2007)*

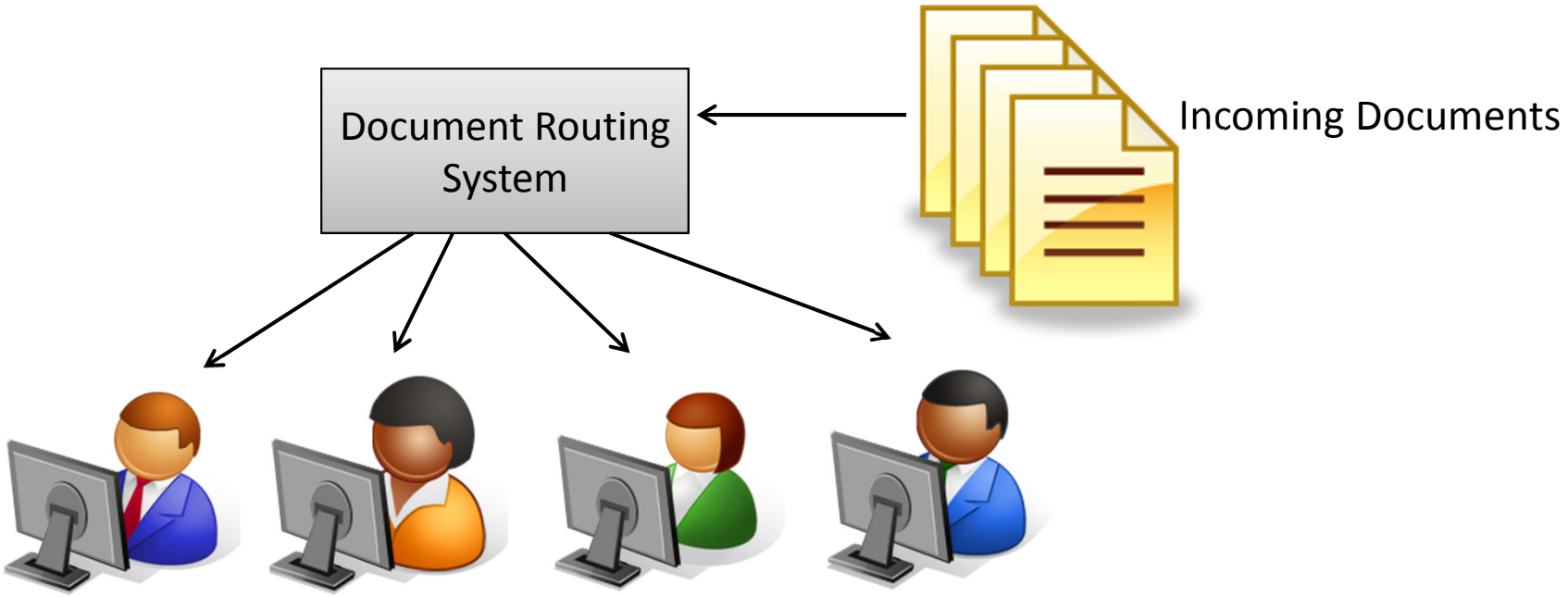
Typical tasks covered in IR

- **Search** (‘ad hoc’ retrieval)
  - Static document collection
  - Dynamic queries
  - Changed dramatically with the rise of the web



Typical tasks covered in IR

- **Filtering**
  - Queries are static
  - The document collection is constantly changing
    - Example: corporate mails routed on predefined queries to different parts of the organizations



Typical other tasks covered in IR

- Clustering
- Categorization
- Recommendation (also kind of a filtering mechanism)
- Summarization
- Question answering
- …

Questions to be answered in this lecture

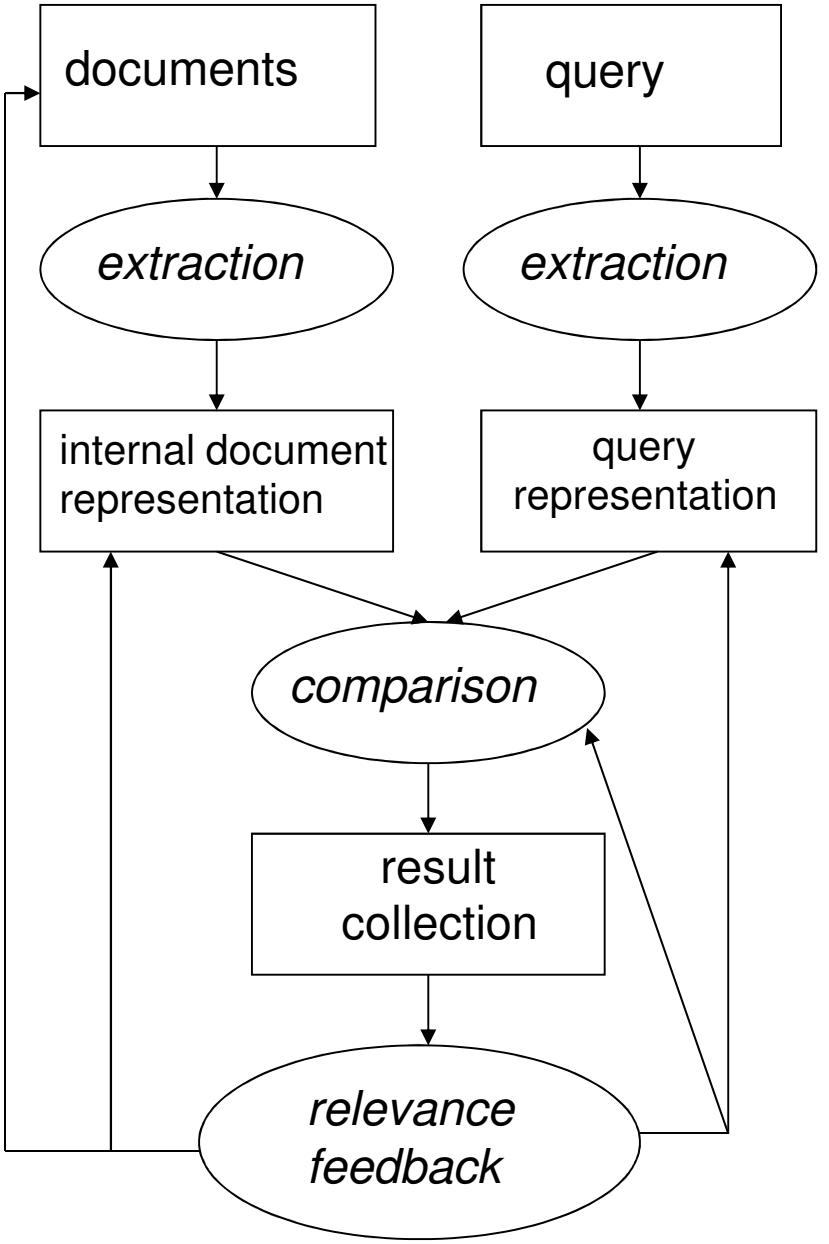
- What is relevance?
- What makes a document relevant?
- How to ensure performance of IR systems?
- How to evaluate the performance of IR systems?

Basic Definitions

Information Retrieval Process

- Information Retrieval Process

1. Extract document and query to an internal representation: *find fixed amount of qualified features, which discribes document (and query) as good as possible*
2. Compare internal representations
3. Collect best results
4. Evaluate relevance, *get feedback from user and modify query automatically -> query iteration*



Large Collections

- Digital society
  - Wide and cheap availability of devices for:
    - Generation,
    - Storage,
    - Processing of digital contents
  - Every N years (N=2 according to given sources, N=5 according to others) the amount of digital information doubles
    - Petabyte --- Exabyte (10^{18} byte, 10^3 Petabyte) – Zetabyte (10^3 Exabyte)
    - often you can read: That's more than in the previous 5,000 years. True?

Unstructured Documents

- **Textual IR**
  - Monolingual – multilingual
  - Structured or unstructured
  - Web pages
  - Scientific papers
  - E-mails
  - Tweets, Blogs, …
  - Newspaper articles
  - Image captions
  - Audio transcript
  - Media annotations (manual or textual)

- **Multimedia IR**
  - Images
  - Graphics
  - Audio (spoken or not spoken)
  - Video
    - !!! All stored in stored in digital form

Information Need

- A document is relevant if it addresses the stated information need, not just because it contains all the word in the query

- E.g., Word in the query:
  - Phython
    - Snake?
    - Programming language?

- E.g., Information need:
  - *I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
  - Query: *wine red white heart attack effective*
  - *Answer: most likely useless*

Information Need and Gaps

- Characterization of information needs is not a simple problem:
  - Sensory Gap
    - Gap between the object in the world and the information in a (computational) description derived from a recording of that scene
  - Semantic Gap
    - Lack of coincidence between the (computational) description of the information and its meaning
- Useful (Relevant), according to the subjective opinion of the user

Information Retrieval is NOT Data Retrieval

- Data Retrieval (RDBMS, XML DB)
  - … retrieving all objects which **satisfy clearly defined conditions** expressed trough a query language.
  - Data has a well defined structure and semantics
  - Formal query languages
    - Regular expression, relation algebra expression, etc.
  - Results are EXACT matches → **errors are not tolerated**
  - No **ranking** w.r.t. the user **information need**
    - Binary retrieval: does not allow the user to control the magnitude of the output
    - For a given query, the system may return:
      - Under-dimensioned output
      - Over-dimensioned output

A Formal Characterization

- An IR model *IRM* can be defined as:

$$IRM =< D, Q, F, R(q_i, d_j) >$$

where

- ***D*** – set of logical views (or representations) for the **Documents** in the collection
- ***Q*** – set of logical views (or representations) for the user's needs. Such representations are called **Queries**
- ***F*** – **Framework** (or strategy) for modeling the document and query representation, and their relationship
- ***R(q_i,d_j)*** – **Ranking function**, associates a real number to a document representation d_j according to a query q_i. Such ranking defines an ordering among the documents with regard to the query q_i

Measures for IR Systems

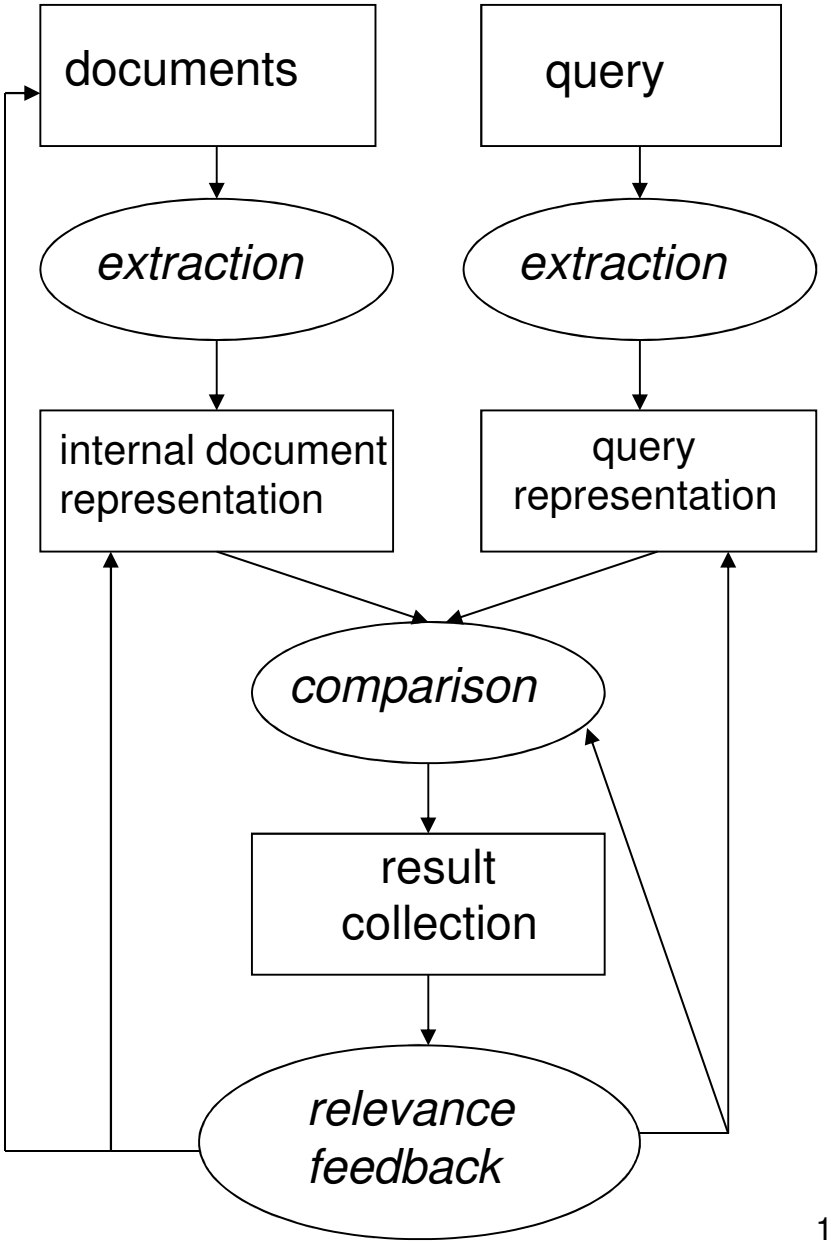
- ***Measurable properties***
  - How fast does it process (index) documents?
    - Number of documents/hour
    - Average document size
  - How fast does it search?
    - Latency as a function of index size
    - Expressiveness of query language
    - Speed on complex queries
- However, the **key** measure is: **user happiness**
  - How to define user happiness?
  - How do we quantify user happiness?

IR Process

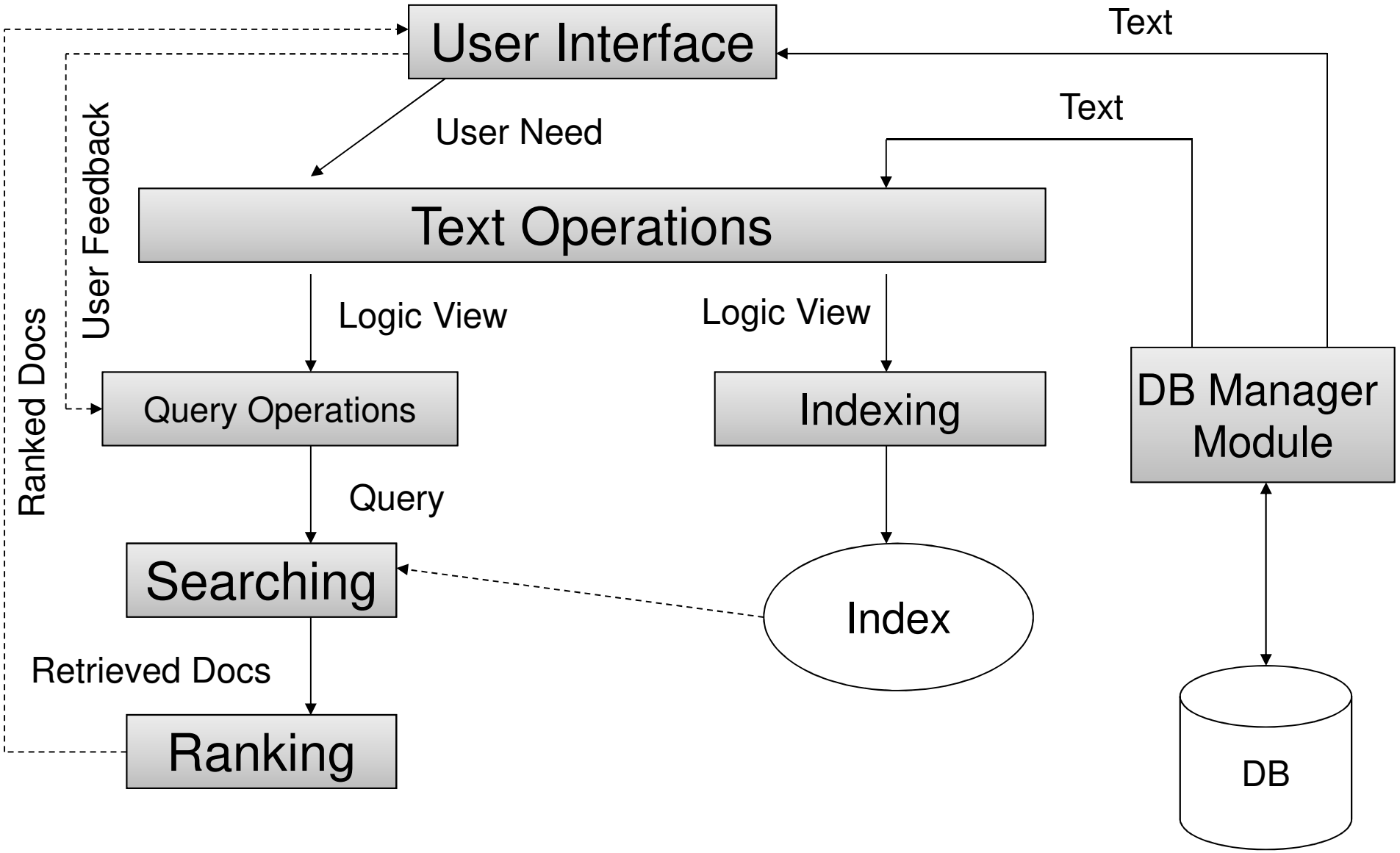
Information Retrieval Process

- Information Retrieval process

1. Extract document and query to an internal representation: *find fixed amount of* <u>*qualified features*</u> *which discribes document (query) as good as possible*
2. Compare internal representations
3. Collect best results
4. Evaluate relevance *get feedback from user and modify query automatically -> query iteration*



High-level View of a Textual IR System



Logic View of Documents

- Documents in a collection are usually represented through a set of **index terms** or **keywords**
  - Index term: any word which appears in the text of a document in the collection
  - Assumption: the semantics of the documents and of the user information need can be naturally expressed through sets of index terms (this is a considerable oversimplification of the problem)
- Keywords are:
  - Extracted directly from the text of the document
  - Specified by a human subject (e.g., tags, comments etc.)

- They provide a ***logic view*** *of the document.*
  - Retrieval systems representing a document by its full set of words use a *full text* logical view (or representation) of the documents.
  - With very large collections, the set of representative have to be reduced by means of TEXT OPERATIONS

Indexing Process

1. Define the text data source
   - usually done by the DB manager, which specifies:
     - Documents
     - Operations to be performed on them
     - Content model (i.e., the content structure and what elements can be retrieved)
2. The content operations transform the original documents and generate a ***logical view*** of them
3. An ***index*** of the text is built on logical view
   - The index allows *fast searching* over large volumes of data. Different index structures might be used, but the most popular one is the ***inverted file***
   - The resources (time and storage space) spent on defining the text database and building the index are amortized by querying the retrieval system many times

Retrieval Process

1. The user first specifies a ***user need***
   - User-level *query (*e.g., keywords); might also be done implicitly (RecSys)
2. The user need is parsed and transformed by the same content operations applied to the indexed contents.
3. *Query operations* provide a system representation for the user need as a system-level query
4. The query is processed to obtain the *retrieved documents.*
   - Fast query processing is made possible by the index structure previously built.
5. The retrieved documents are ranked according to a *likelihood* of relevance.
6. The user then examines the set of ranked documents in the search for useful information.
   - he might pinpoint a subset of the documents seen as definitely of interest and initiate a user feedback cycle.

Credits

- Slides partly adapted from
  - Eva Zangerle, DBIS Innsbruck (2014/15)
  - Stefano Ceri, Alessandro Bozzon , Marco Brambilla,Emanuele Della Valle, Piero Fraternali, Silvia Quarteroni: Web Information Retrieval
  - Dietmar Jannach, Markus Zanker, Alexander Felfernig, Gerhard Friedrich: Recommender Systems – An Introduction
  - Günther Specht, DBIS Innsbruck (former lectures)