

Chapter One: Introduction to ISR

Outline

- IR and IR systems
- Data retrieval versus information retrieval
- IR and the retrieval process
- Basic structure of an IR system

What is information?
What are the Sources of information?
Why we use information?

- ❖ **Information**: data that are processed to be useful; provides answers to "who", "what", "where", and "when" questions.
- ❖ Information is data that has been given **meaning** by way of relational connection. This "meaning" can be useful, but does not have to be.
- ❖ In computer parlance, a relational database makes information from the data stored within it.

❖ Information Storage

- The computer stores **different types of information** in different ways, depending on what the information is, how much storage space it requires, and how quickly it needs to be accessed.

Sources of information

- Information is organized into (a large number of) documents
 - Large collections of documents available from various **sources**: books, magazines, newspapers, journal articles, conference papers, digital libraries, Web pages, etc.
- **Example:**
 - Google Web Search Engine claims to index over 3 billion pages
 - AltaVista, www.altavista.com, covers over 250 million Web pages.
 - It performs more than 40 million search queries each day in more than 25 languages

Retrieval ?

- Information retrieval is the process of **searching** for **relevant** documents from **unstructured large** corpus that satisfy information **need** of users.
 - It is a tool that finds and selects from a collection of items a subset that serves the user's purpose.



- ❖ 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

General Goal of Information Retrieval

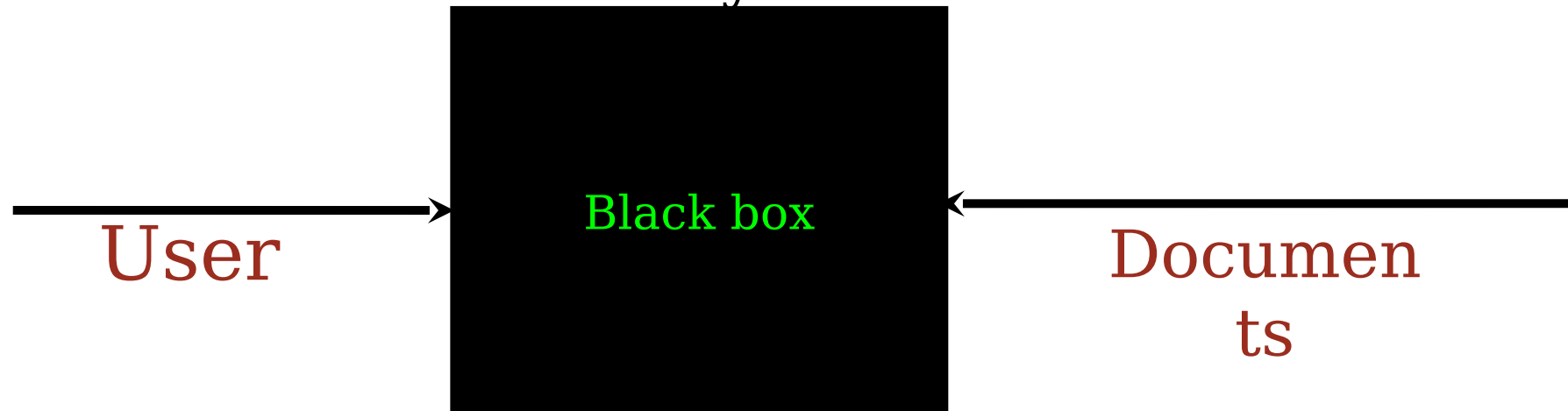
- To help users find **useful information** based on their information needs (with a minimum effort) despite
 - Increasing complexity of Information
 - Changing needs of user
- Provide **immediate random access** to the document collection.
- Retrieval systems, such as Google, Yahoo, are developed with this aim.

Examples of IR systems

- Much IR research focuses more specifically on text retrieval. But there are many other interesting areas:
 - ❖ Cross-language retrieval, text summarization, information filtering, Question-answering, content-based multimedia (audio, Image and Video) retrieval.
- Text-based (Lexis-Nexis, Google, FAST): Search by keywords. Limited search using queries in natural language.
- Multimedia (QBIC, WebSeek, SaFe): Search by visual appearance (shapes, colors,...).
- Question answering systems (AskJeeves, Answerbus): Search in (restricted) natural language
- Cross language vs. multilingual information retrieval

Information Retrieval serve as Bridge

- An Information Retrieval System serves as a bridge between the world of authors and the world of readers/users,
 - That is, writers present a set of ideas in a document using a set of concepts.
 - Then Users seek the IR system for relevant documents that satisfy their information need.



Information Retrieval vs. Data Retrieval

- Emphasis of IR is on the retrieval of information, rather than on the retrieval of data
- **Data retrieval**
 - ◆ Consists mainly of determining which documents contain a set of keywords in the user query (which is not enough to satisfy the user information need)
 - ◆ Aims at retrieving all objects that satisfy well defined semantics
 - ◆ a single erroneous object among a thousand retrieved objects implies failure
- **Information retrieval**
 - ◆ It is concerned with retrieving information about a subject or topic than retrieving data which satisfies a given query
 - ◆ Semantics is frequently loose: the retrieved objects might be inaccurate
 - ◆ Small errors are tolerated

Information Retrieval vs. Data Retrieval

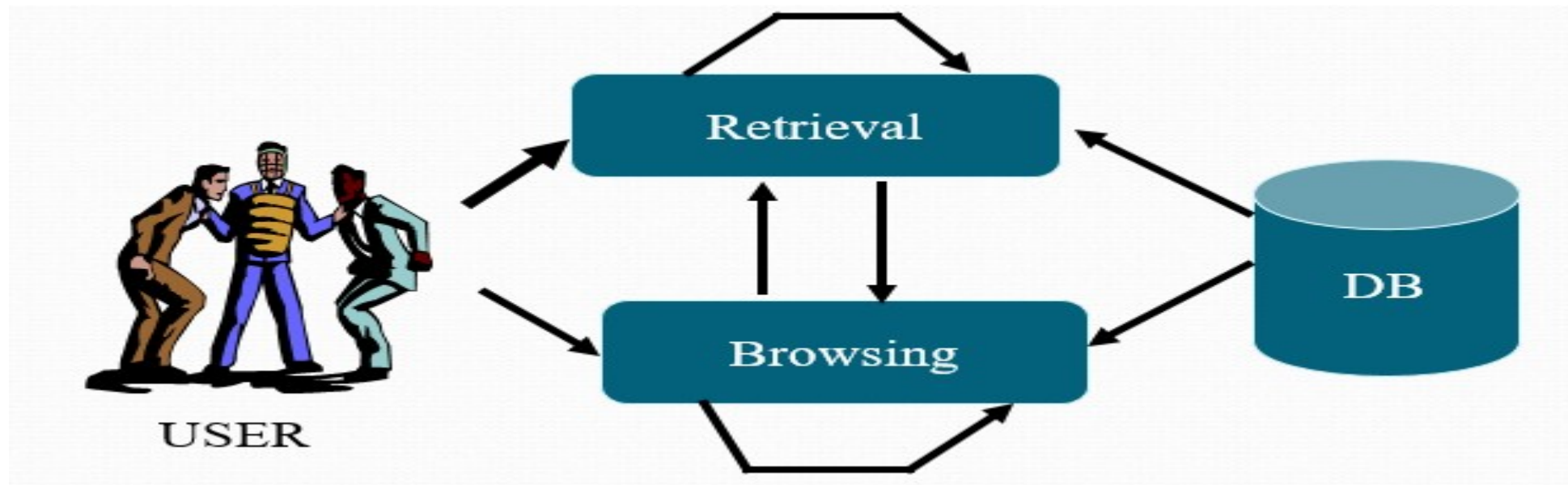
	Data Retrieval	Information Retrieval
Data organization	Structured (Clear Semantics: Name, age,...)	Unstructured (No fields (other than text))
Query Language	Artificial (defined, SQL)	Free text (“natural language”), Boolean
Query specification	Complete	Incomplete
Items wanted	Exact Matching	Partial & Best matching, Relevant
Accuracy	100 % (results are <i>always</i> “correct”)	< 50 %
Error response	Sensitive	Insensitive

Basic Concepts in Information Retrieval:

(i) User Task and (ii) Logical View of documents

The User Task:

❖ Two user task - **retrieval** and **browsing**



The User Task

Retrieval

- ❖ It is the process of retrieving information whereby the main objective is clearly defined from the onset of searching process.
- ❖ The user of a retrieval system has to translate his information need into a query in the language provided by the system.
- ❖ In this context (i.e. by specifying a set of words), the user searches for useful information executing a retrieval task.
- ❖ English Language Statement:
I want a book by J. K Rowling titled The Chamber of Secrets

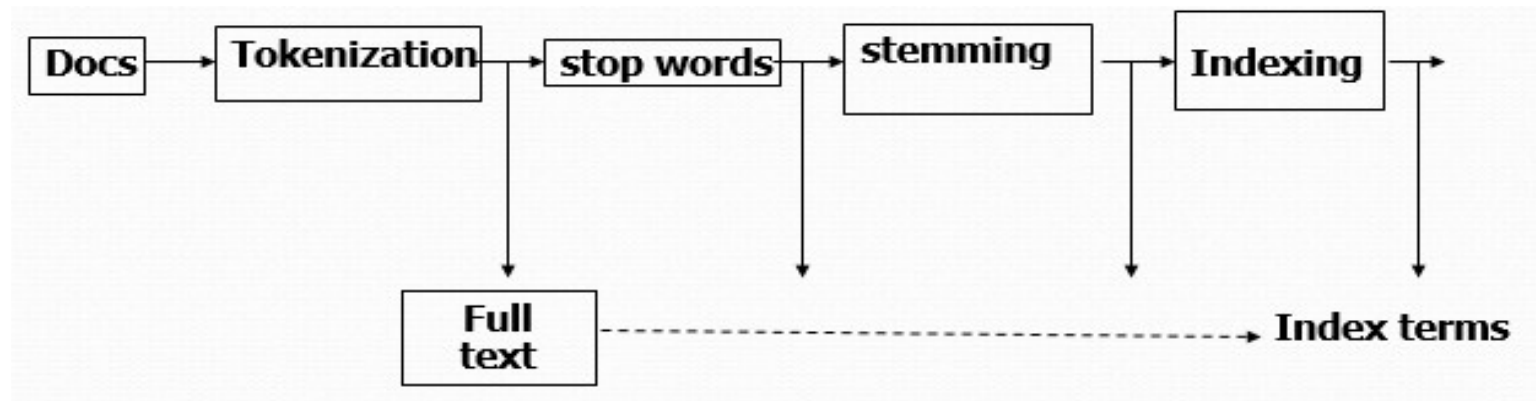
The User Task

Browsing

- It is the process of retrieving information, whereby the main objective is not **clearly defined** from the beginning and whose purpose might change during the interaction with the system.
- E.g. User might search for documents about 'car racing'. Meanwhile he might find interesting documents about 'car manufacturers'. While reading about car manufacturers in Addis, he might turn his attention to a document providing 'direction to Addis', and from this to documents which cover 'Tourism in Ethiopia'.
- In this context, user is said to be browsing in the collection and not searching, since a user may have an interest glancing around

Logical View of Documents

- ❖ Documents in a collection are frequently represented by a set of index terms or keywords
- ❖ Such keywords are mostly extracted directly from the text of the document
- ❖ These representative keywords provide a logical view of the document



- ❖ Document representation viewed as a continuum, in which logical view of documents might shift from full text to index terms.

Logical view of documents

- If full text :
 - Each word in the text is a keyword
 - Most complex form
 - Expensive
 - If full text is too large, the set of representative keywords can be reduced through transformation process called text operation
- It reduce the complexity of the document representation and allow moving the logical view from that of a full text to a set of index terms

Typical IR Task

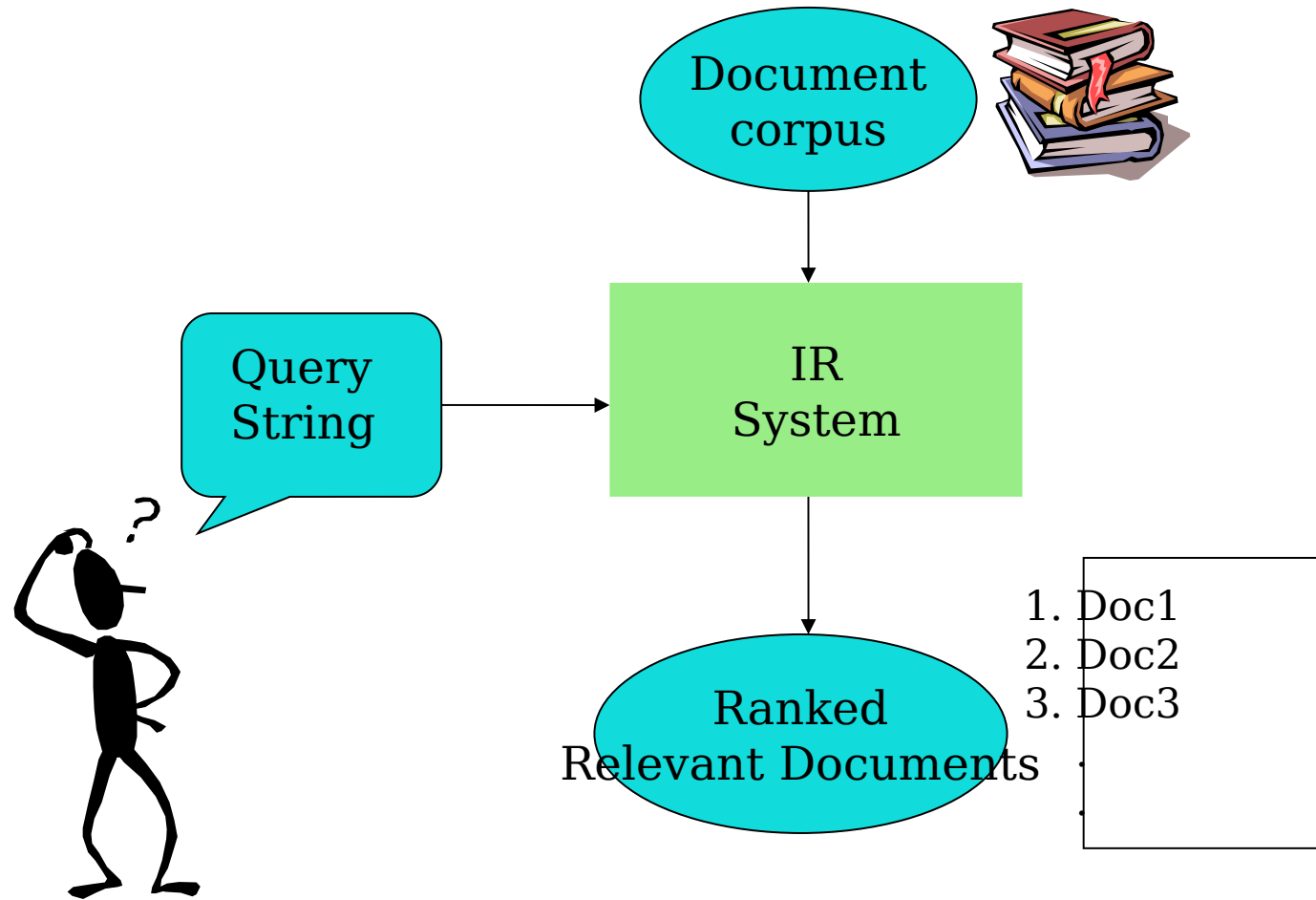
➤ Given:

- ❖ A corpus of textual natural-language documents.
- ❖ A user query in the form of a textual string.

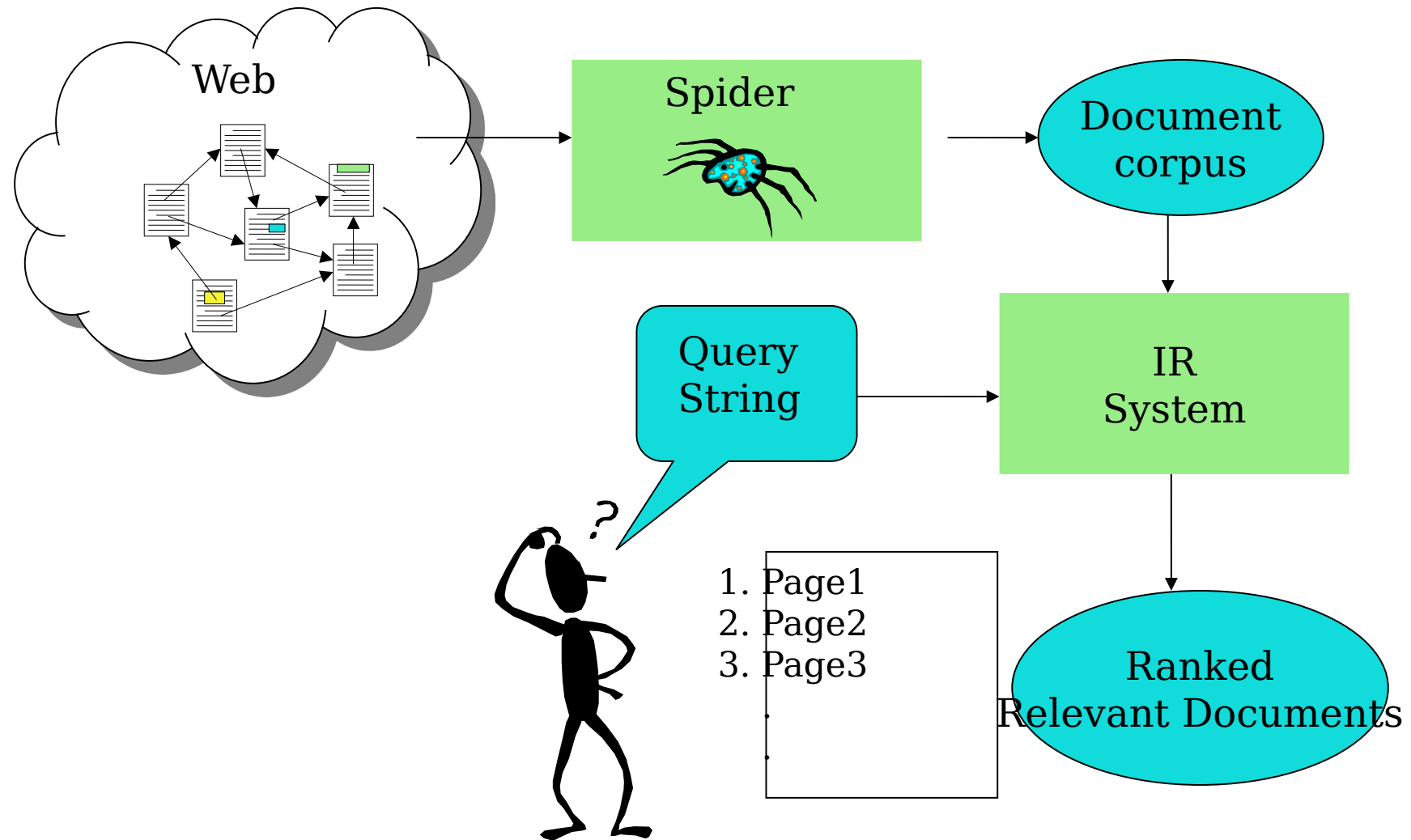
❖ Find:

- ❖ A ranked set of documents that are relevant to the query.

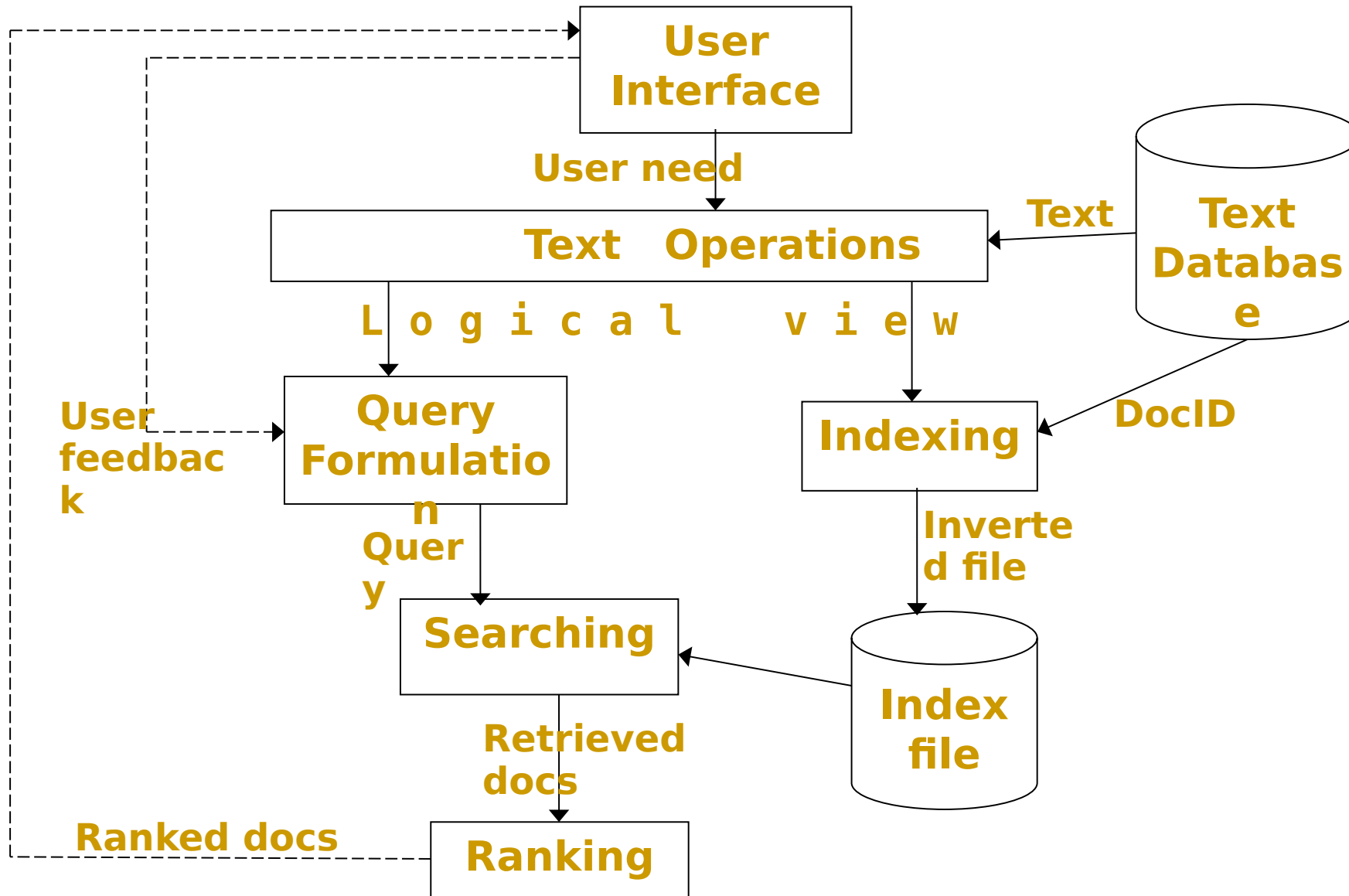
Typical IR System Architecture



IR System vs. Web Search System



The Retrieval Process



The Retrieval Process

- It is necessary to **define the text database** before any of the retrieval processes are initiated.
- The **text operations** transform the original documents & the information needs and generate a logical view of them
- Once the logical view of the documents is defined, the database module **builds an index** of the text
 - An index is a critical data structure
 - It allows fast searching over large volumes of data
- Different index structures might be used, but the most popular one is the inverted file.
- Given the document database is indexed, the **retrieval process can be initiated**

The Retrieval Process

- The user first **specifies a user need** which is then **parsed & transformed** by the same text operation applied to the text
 - Next the **query operations is applied** before the actual query, which provides a system representation for the user need, is generated
- The query is then **processed** to retrieve documents
 - Before the retrieved documents are sent to the user, the retrieved documents are **ranked** according to the likelihood of relevance
- The user then **examines** the set of ranked documents in the search for useful information. Two choices for the user:
 - (i) reformulate query, run on entire collection or (ii) reformulate query, run on result set
- At this point, the user might pinpoint a subset of the documents seen as definitely of interest & initiate a user **feedback** cycle
 - In such a cycle the system uses the documents selected by the user

Issues that arise in IR

- Text document representation
 - what makes a “good” representation?
 - how is a representation generated from text?
 - what are the retrievable objects & how are they organized?
- Information need representation
 - what is an appropriate query language?
 - how can interactive query formulation & refinement be supported?
- Comparing representations
 - what is a “good” similarity measure & retrieval model?
 - how is uncertainty represented?
- Evaluating effectiveness of retrieval
 - what are good metrics?
 - what constitutes a good experimental test bed?

Information Retrieval Research areas

- Much of IR research focuses more specifically on *text retrieval*. But there are many other interesting areas:
 - ❖ **Audio retrieval**, which deals with searching for speech or music file.
 - ❖ **Cross-language retrieval**, which uses a query in one language (say English) and finds documents in other languages (say Afaan Oromo and Amharic).
 - ❖ **Question-answering IR systems**, which retrieve answers from a body of text.
 - ✓ For example, the question Who won the 1997 World Series? finds a 1997 headline World Series: Marlins are champions.
 - ❖ **Image retrieval**, which finds images on a given topic or images that contain a given shape or color.

Is IR just document retrieval?

- Cross-language information retrieval
- Information extraction
- Question answering
- Document Summarization
- Text classification
- Multimedia information retrieval
- Multi-database searching
- Document provenance
- Agents (information filtering, tracking, routing)
- Recommender systems
- Text mining

End of Chapter-1