



A T M E
College of Engineering



Information Retrieval

Module 1



- Introduction
- Information versus Data Retrieval
- IR: Past, present, and future
- Basic concepts: The retrieval process
- Logical view of documents
- Modeling: A Taxonomy of IR models
- Ad-hoc retrieval and filtering
- Classic IR models
- Set theoretic
- Algebraic
- Probabilistic IR models
- Models for browsing



- Information retrieval (IR) deals with the representation, storage, organization and access to information items.
- The representation and organization of the information items should provide the user with easy access to the information in which he is interested.
- Unfortunately, characterization of the user information need is not a simple problem.
- Web search engines are the most visible IR applications.
- The user must first translate the request information into a query which can be processed by the search engine.



- An information retrieval process begins when a user enters a query into the system.
- Queries are formal statements of information needs, for example search strings in web search engines.
- Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user.
- The emphasis is on the retrieval of information as opposed to the retrieval of data.
- Data is a raw fact and information is the processed data.

Information versus Data Retrieval

Information Retrieval	Data Retrieval
Retrieves <i>information</i> about a subject	Determines the <i>keywords</i> in the user query and retrieves data
Small errors are likely to go unnoticed.	Single erroneous object means total failure
Deals with natural language text	Deals with relational database
Not always well structured and is semantically ambiguous.	Has a well defined structure and semantics
Does not provides solution to the user of a database system	Provides solution to the user of a database system
Extracts syntactic and semantic information from the document text and use this information to match the user information need	Does not solve the problem of retrieving information about a subject
The primary goal is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible.	The primary goal is to retrieve the data according to the user query



IR: Past, present, and future

- A typical example of information retrieval is the table of contents of a book
- Since the volume of information eventually grew beyond a few books, it became necessary to build specialized data structures to ensure faster access to the stored information.
- An old and popular data structure for faster information retrieval is a collection of selected words or concepts with which are associated pointers to the related information is the **index**.
- In one form or another, indexes are at the core of every modern information retrieval system.
- They provide faster access to the data and allow the query processing task to be speeded up.



- Two different views of the IR problem: a computer-centered one and a human-centered one.
- In the computer-centered view, the IR problem consists mainly of building up efficient indexes, processing user queries with high performance, and developing ranking algorithms which improve the 'quality' of the answer set.
- In the human-centered view, the IR problem consists mainly of studying the behavior of the user, of understanding his main needs, and of determining how such understanding affects the organization and operation of the retrieval system



systems for retrieving information.

- In the first generation, the systems allowed searches based on author name and title.
- In the second generation, increased search functionality was added which allowed searching by subject headings, by keywords.
- In the third generation, which is currently being deployed, the focus is on improved graphical interfaces, electronic forms, hypertext features, and open system architectures.



A T M E
College of Engineering



- Three dramatic and fundamental changes have occurred due to the advances in modern computer technology and the boom of the Web.
 - First, it became a lot cheaper to have access to various sources of information.
 - Second, the advances in all kinds of digital communication provided greater access to networks.
 - Third, the freedom to post whatever information someone judges useful has greatly contributed to the popularity of the Web.
- Fundamentally, low cost, greater access, and publishing freedom have allowed people to use the Web as a highly *interactive* medium.
- Such interactivity allows people to exchange messages, photos, documents, software, videos, and to 'chat' in a convenient and low cost fashion.



A T M E
College of Engineering



Practical Issues

- Security
- Privacy
- Copyright



- To describe the retrieval process, we use a simple and generic software architecture .
- First of all, before the retrieval process can even be initiated, it is necessary to define the text database.
- This is usually done by the manager of the database, which specifies the following:
 - (a) The documents to be used
 - (b) The operations to be performed on the text
 - (c) The text model (i.e., the text structure and what elements can be retrieved). The text operations transform the original documents and generate a logical view of them.
- Once the logical view of the documents is defined, the database manager (using the DB Manager Module) builds an index of the text.



- Given that the document database is indexed, the retrieval process can be initiated.
- The user first specifies a user need which is then parsed and transformed by the same text operations applied to the text.
- Then, query operations might be applied before the actual query, which provides a system representation for the user need, is generated.
- The query is then processed to obtain the retrieved documents.
- Fast query processing is made possible by the index structure previously built.
- Before being sent to the user, the retrieved documents are ranked according to a likelihood of relevance.
- The user then examines the set of ranked documents in the search for useful information.

- In such a cycle, the system uses the documents selected by the user to change the query formulation.

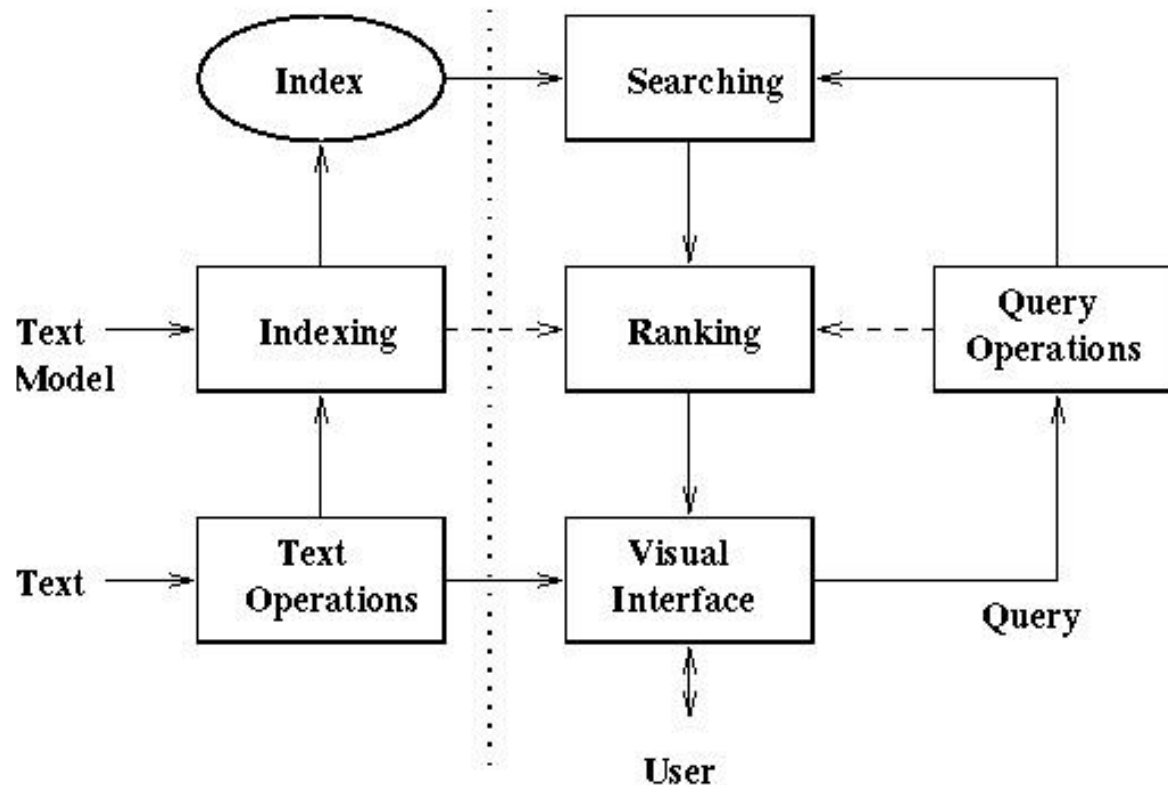


Figure: The process of retrieving information (the numbers beside each box indicate the chapters that cover the corresponding topic).



- Due to historical reasons, 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.
- No matter whether these representative keywords are derived automatically or generated by a specialist, they provide a logical view of the document.
- Modern computers are making it possible to represent a document by its full set of words. In this case, we say that the retrieval system adopts a *full text* logical view of the documents.



A T M E
College of Engineering



- This can be accomplished through the elimination of *stopwords* (such as articles and connectives), the use of *stemming* (which reduces distinct words to their common grammatical root), and the identification of *noun groups* (which eliminates adjectives, adverbs, and verbs).
- These operations are called *text operations* (or transformations).
- Text operations reduce the complexity of the document representation and allow moving the logical view from that of a full text to that of a set of *index terms*.

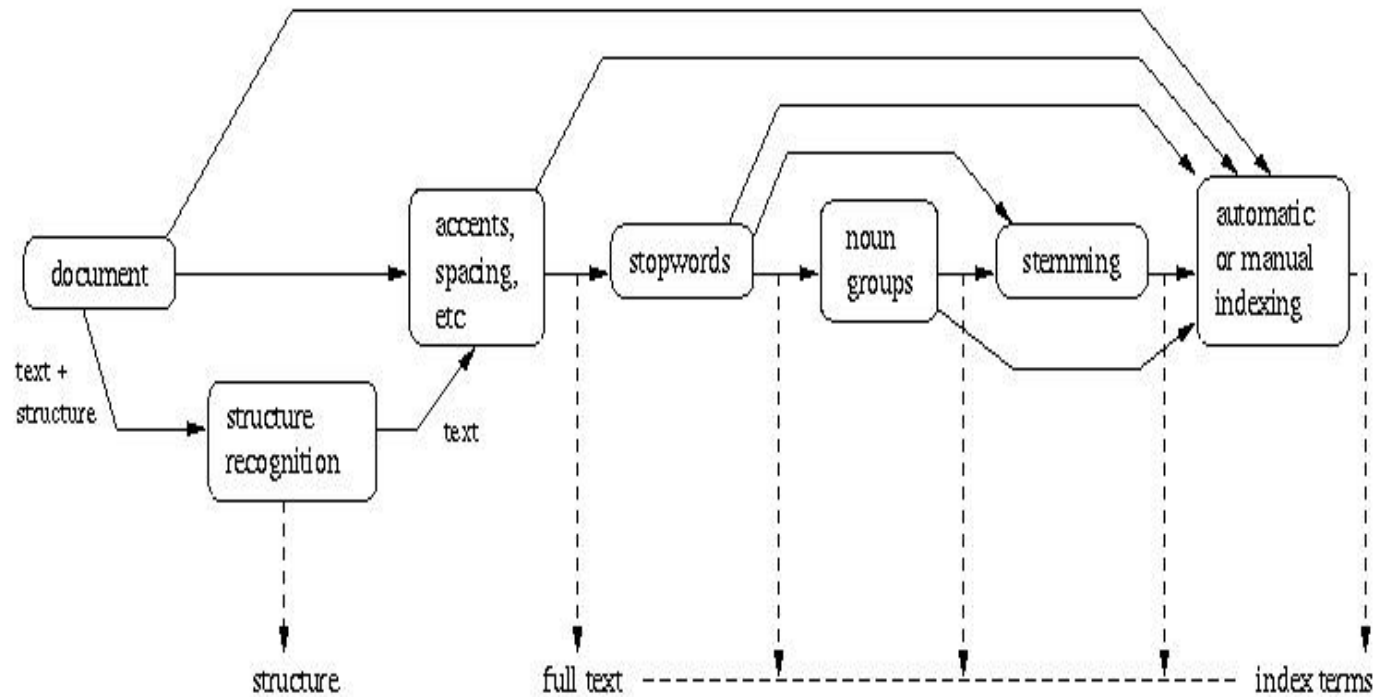


Figure: Logical view of a document: from full text to a set of index terms.