

Chapter One

Introduction to IR System

1.1 Introduction

As the amount of online textual information (e.g., web pages, email, news articles, office documents, and scientific literature) grows explosively, it is increasingly important to develop tools to help us manage and exploit the huge amount of information. Web search engines, such as Google, Yahoo!, and MSN, are good examples of such tools, and they are now an essential part of everyone's life.

As the underlying science of search engines, information retrieval (IR) has been studied since several decades ago, but the huge impact of the research results of the information retrieval community had not appeared until the birth of the Web. Now information retrieval has become a very active research area, attracting more and more attention recently. The purpose of this course is two-fold: (1) introduce the foundational concepts, principles, and techniques of IR and review a representative set of frontier topics. (2) Discuss the general methodology and specific strategies for doing research. IR deals with the representation, storage, organization of 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/she is interested. Focus is on the user information need.

What types of information?

- Text (Documents and portions thereof)
- XML and structured documents
- Images
- Audio (sound effects, songs, etc.)
- Video
- Source code
- Applications/Web services

Types of Information Needs:

- Retrospective

- “Searching the past”
- Different queries posed against a static collection
- Time invariant
- Prospective
 - “Searching the future”
 - Static query posed against a dynamic collection
 - Time dependent

Retrospective Searches:

- *Ad hoc* retrieval: find documents “about this”
 - Identify positive accomplishments of the Hubble telescope since it was launched in 1991.
 - Compile a list of mammals that are considered to be endangered, identify their habitat and, if possible, specify what threatens them.
- Known item search
 - Find Jimmy Lin’s homepage.
 - What’s the ISBN number of “Modern Information Retrieval”?
- Directed exploration
 - Who makes the best chocolates?
 - What video conferencing systems exist for digital reference desk services?
- Question answering
 - “Factoid”: Who discovered Oxygen?
When did Hawaii become a state?
Where is Ayer’s Rock located?
What team won the World Series in 1992?
 - “List”: What countries export oil?
Name U.S. cities that have a “Shubert” theater
 - “Definition”: Who is Aaron Copland?
What is a quasar?

Prospective “Searches”

- Filtering
 - Make a binary decision about each incoming document-Spam or not spam?

- Routing
 - Sort incoming documents into different bins?
 - Categorize news headlines: World? Nation? Metro? Sports?

1.2 Information versus data retrieval

- Data retrieval: which documents contain the keywords in the user query?
 - Clearly defined conditions
 - such as regular expression or relation algebra expression
 - A single erroneous object means total failure
 - Well-defined structure and semantics
 - A single erroneous object among a thousand objects means total failure.
 - Example of data retrieval system is a relational database
- Information retrieval: information about a subject or topic which satisfies a given query.
 - Not well-structured or semantically ambiguous
 - Small errors are unnoticed.
 - Small errors are allowed

The basic difference between data and information retrieval are presented in the table below

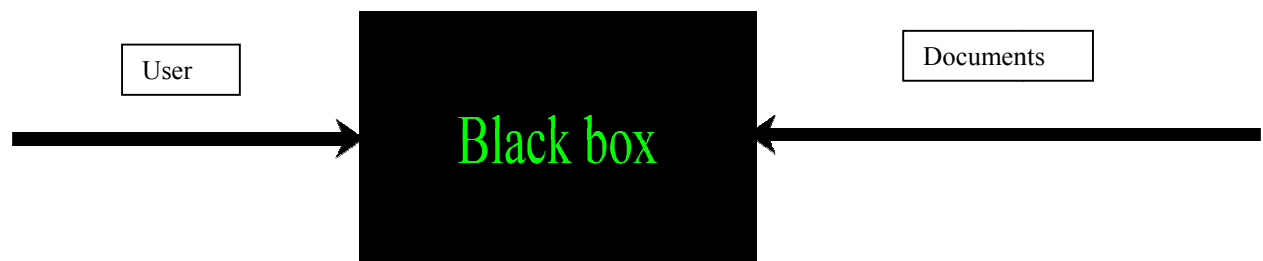
	Data Retrieval	Info Retrieval
Results we get (Matching)	Exact match	Partial match, best match
Data organization	Structured (Clear Semantics: Name, age,...)	Unstructured (No fields (other than text))
Model	Deterministic	Probabilistic
Accuracy	100 % (results are <i>always</i> “correct”)	< 50 %
Query language	Artificial (SQL)	Free text (“natural language”),
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant

Error response	Sensitive	Insensitive
Interaction with system	One-shot queries	Interaction is important

In general, Data retrieval (DR) and information retrieval (IR) have traditionally occupied two distinct niches in the world of information systems. DR systems effectively store and query structured data, but lack the flexibility of IR, i.e., the ability to retrieve results which only partially match a given query. IR, on the other hand, is quite useful for retrieving partial matches, but lacks the completed query specification on semantically unambiguous data of DR systems. Due to these drawbacks, we propose an approach to combine the two systems using pre-defined *word similarities* to determine the correlation between a keyword query (commonly used in IR) and data records stored in the inner framework of a standard RDBMS. Our integrated approach is flexible, context-free, and can be used on a wide variety of RDBs. Experimental results show that RDBMSs using our word-similarity matching approach achieve high mean average precision in retrieving relevant answers, besides exact matches, to a keyword query, which is a significant enhancement of query processing in RDBMSs.

1.3 Structure of an IR System

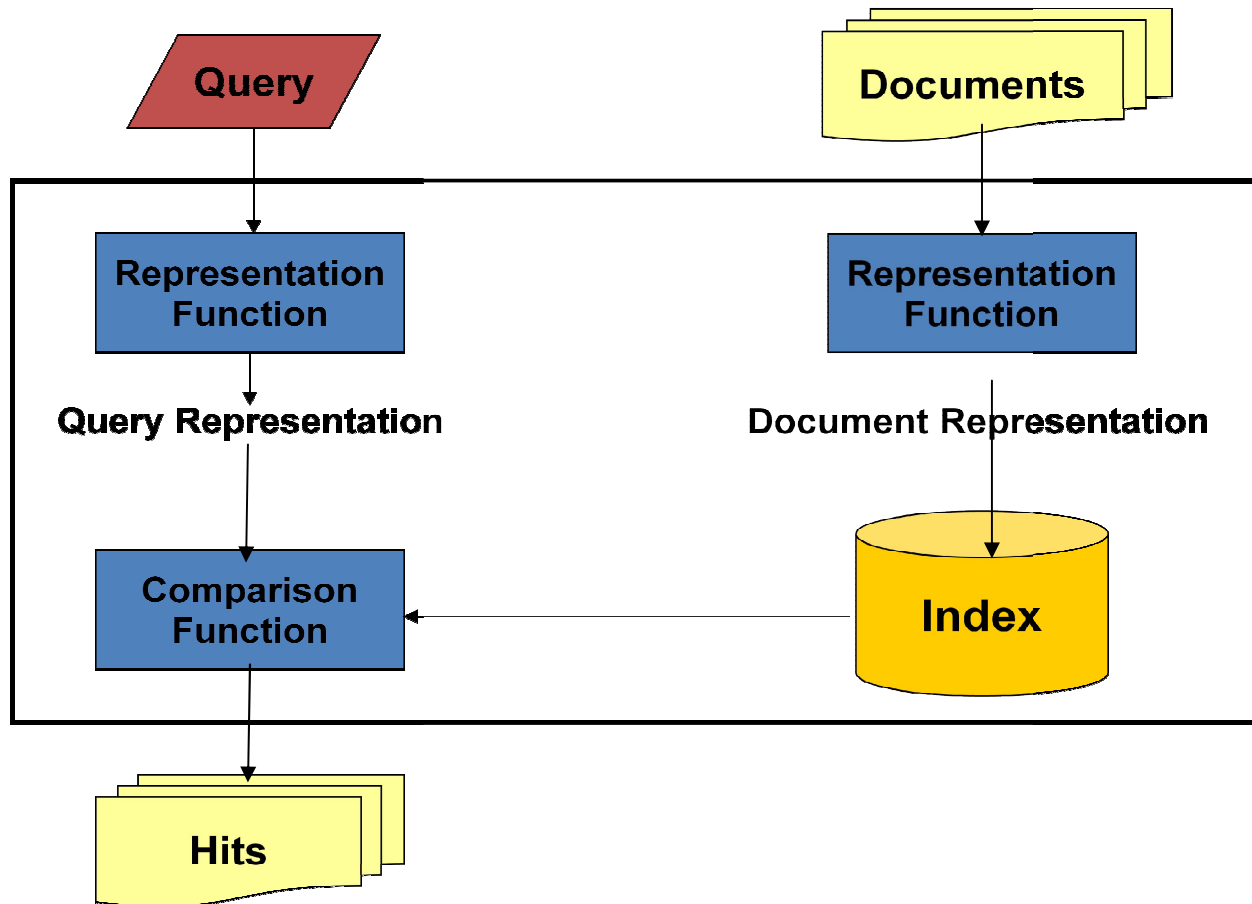
- An Information Retrieval System serves as a bridge between the world of authors and the world of readers/users,
 - That is, writers present a set of ideas in a document using a set of concepts. Then Users seek the IR system for relevant documents that satisfy their information need.
 - The user of t a retrieval system has to translate his information need into a query in the language provided by the system. With an information retrieval system, this normally implies specifying a set of words which convey the semantics of information need.



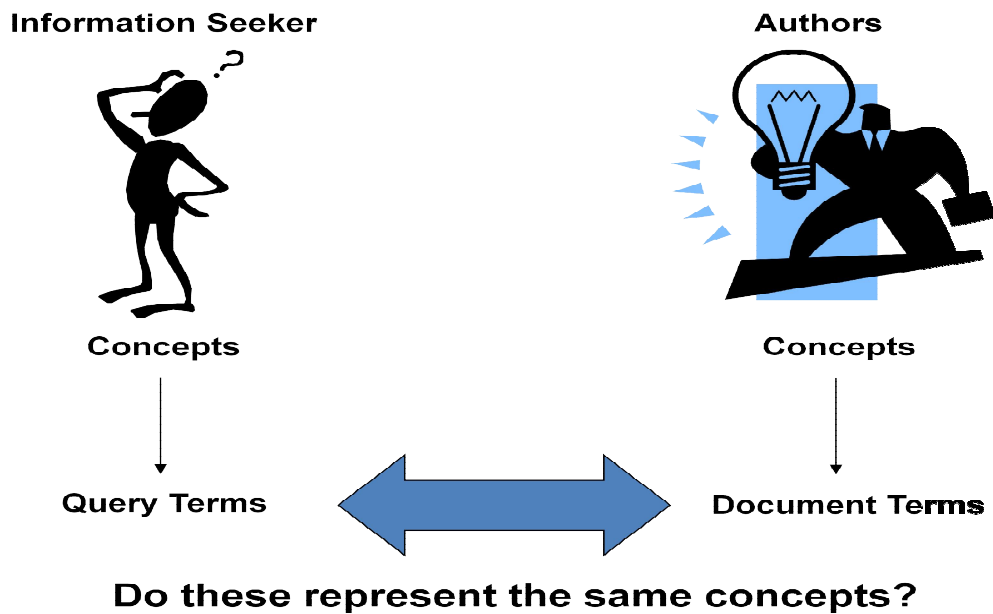
- What is in the Black Box?

- The black box is the processing part of the information retrieval system
- This course gives a highlight of how IR system works ?

Inside The IR Black Box



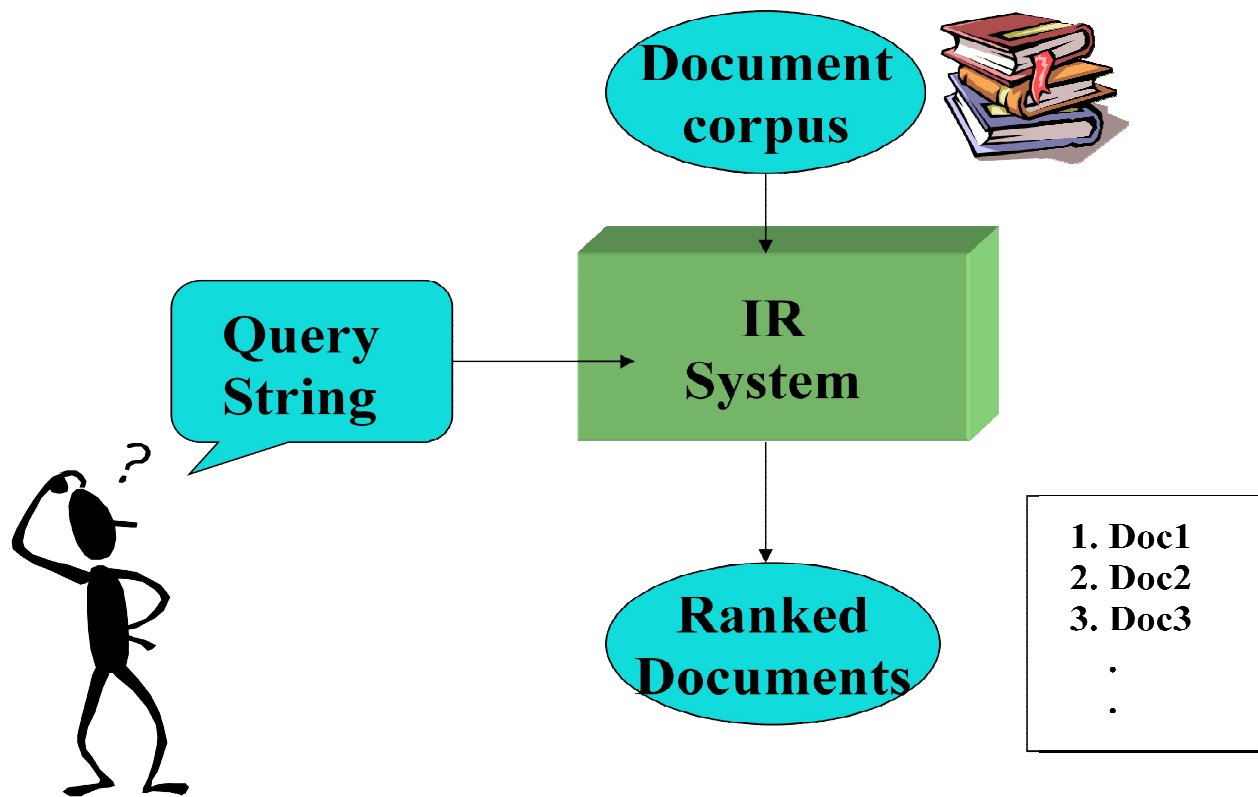
The Central Problem in IR



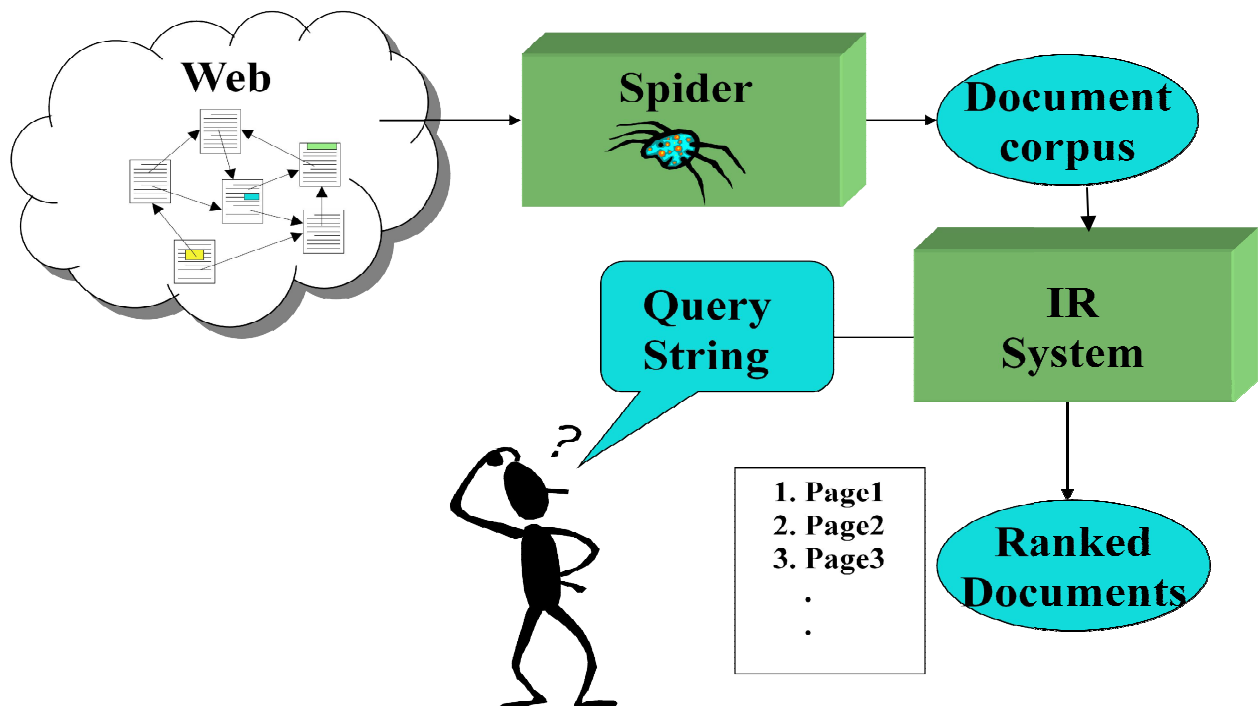
Typical IR Task

- Given:
 - A corpus of document collections (text, image, video, audio) published by various authors.
 - A user information need in the form of a query.
- An IR system searches for:
 - A ranked set of documents that are relevant to satisfy information need of a user.

Typical IR System Architecture



Web Search System

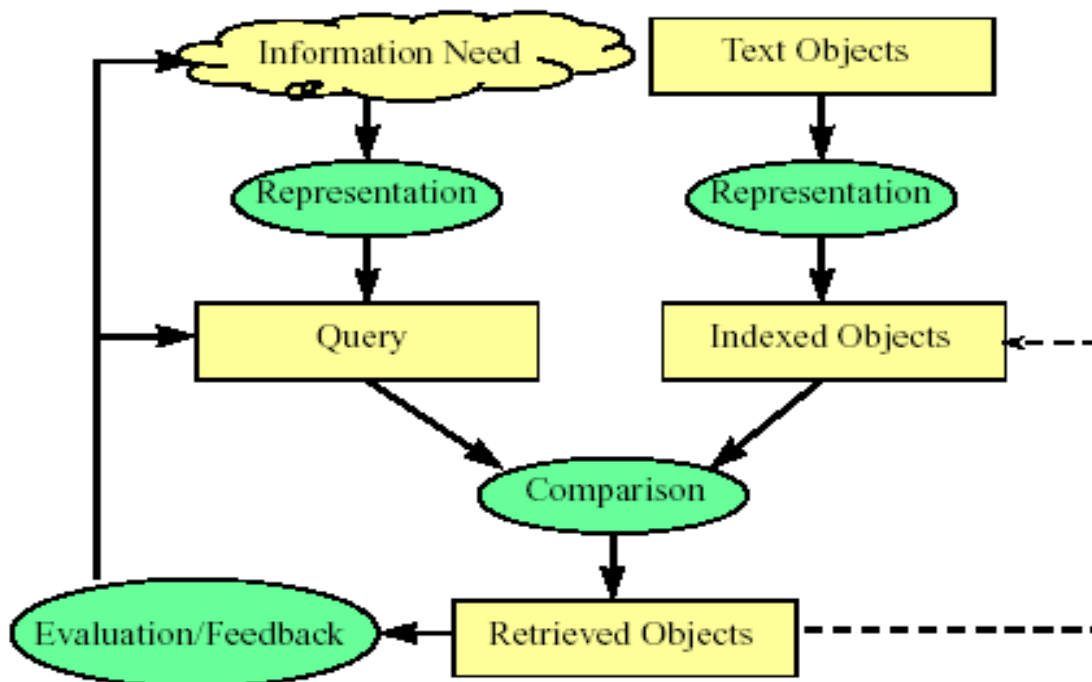


What is Information Retrieval?

Information retrieval deals with representation, storage, organization of, and access to relevant information that satisfy users information need.

- Features of a good information retrieval system:
 - Representation
 - Storage
 - Organization
 - Access
 - Evaluation

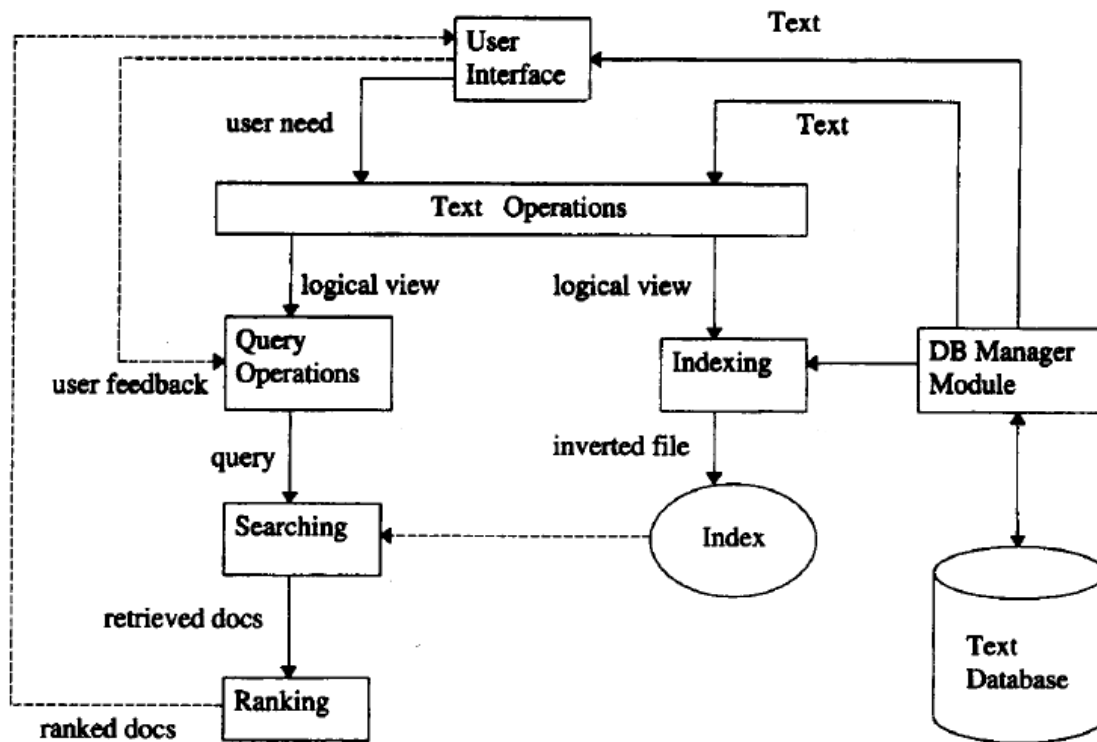
Overview of the Retrieval Process



Main ingredients of IR Process

- There are three main ingredients
 - Texts or documents
 - Queries
 - The process of evaluation
- Texts/Documents Representation
 - Obtaining a representation of the text

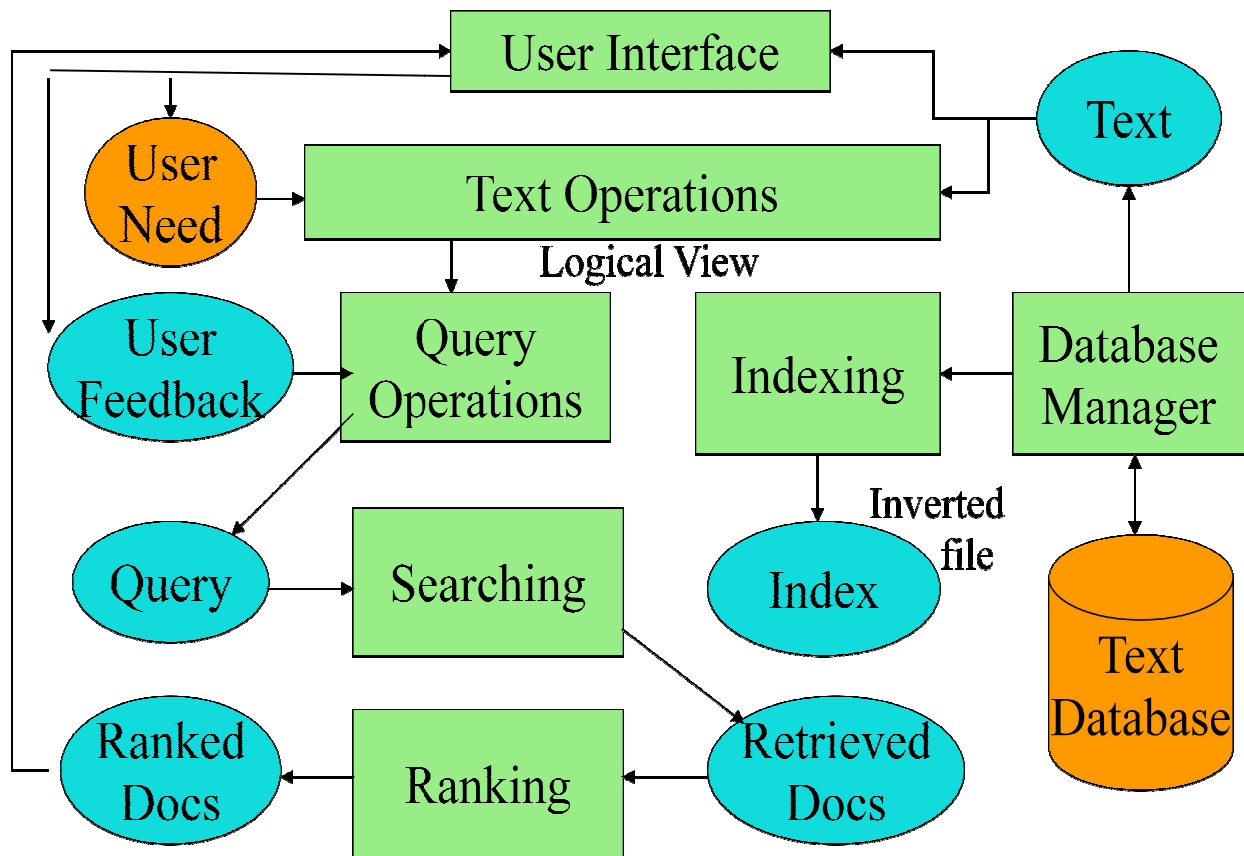
- The representation is achieved by creating an abbreviated form of the text, known as a text surrogate
- A typical surrogate would consist of a set of terms or keywords or descriptors



Query representation

- The query has arisen as a result of an information need on the part of the user
 - It is a representation of the information need and must be expressed in a language understood by the system
 - Due to the inherent difficulty of accurately representing the information need, the query in IR system is always regarded as approximate and imperfect
- The evaluation process
 - involves a comparison of the text actually retrieved vs. the user expected to retrieve
 - This is related with the process of measuring the effectiveness of the retrieval operation (recall, precision, ...)

IR System Architecture



Effectiveness and efficiency of IRS

Much of the research and development in information retrieval is aimed at improving the effectiveness and efficiency of retrieval system. Efficiency is usually measured in terms of the computer resources used such as core, backing store, and C.P.U. time. It is difficult to measure efficiency in a machine independent way. In any case, it should be measured in conjunction with effectiveness to obtain some idea of the benefit in terms of unit cost.

Effectiveness is commonly measured in terms of precision and recall.

- Precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved, and
- Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents (both retrieved and not retrieved).