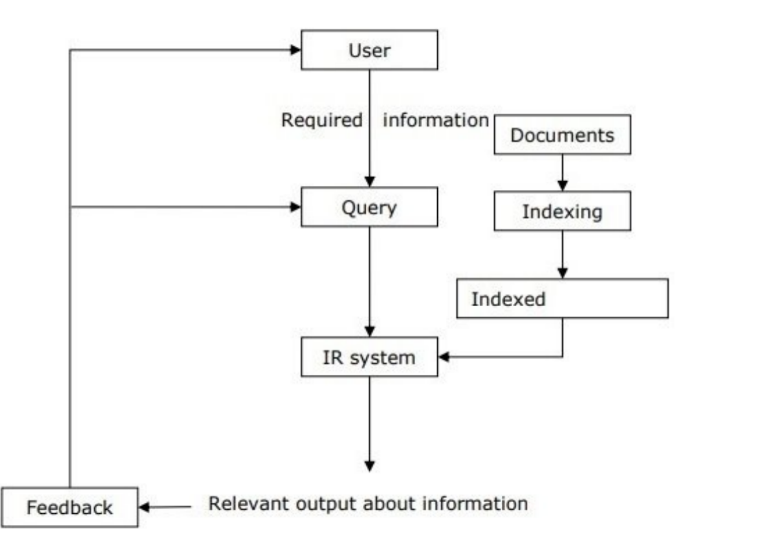
* Information Retrieval is the activity of obtaining material that can usually be documented on an unstructured nature i.e. usually text which satisfies an information need from within large collections which is stored on computers. For example, Information Retrieval can be when a user enters a query into the system.
* The system assists users in finding the information they require but it does not explicitly return the answers of the questions. It informs the existence and location of documents that might consist of the required information.
* The documents that satisfy user’s requirement are called relevant documents. A perfect IR system will retrieve only relevant documents.
* A set of keywords are required to search. Keywords are what people are searching for in search engines. These keywords summarize the description of the information.



What is an IR Model?

An Information Retrieval (IR) model selects and ranks the document that is required by the user or the user has asked for in the form of a query. The documents and the queries are represented in a similar manner, so that document selection and ranking can be formalized by a matching function that returns a retrieval status value (RSV) for each document in the collection. Many of the Information Retrieval systems represent document contents by a set of descriptors, called terms, belonging to a vocabulary V. An IR model determines the query-document matching function according to four main approaches:

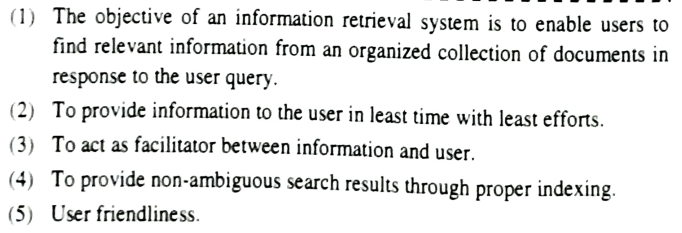
| **Information Retrieval** | **Data Retrieval** |
| --- | --- |
| The software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information. | Data retrieval deals with obtaining data from a database management system such as ODBMS. It is A process of identifying and retrieving the data from the database, based on the query provided by user or application. |
| Retrieves information about a subject. | Determines the keywords in the user query and retrieves the data. |
| Small errors are likely to go unnoticed. | A single error object means total failure. |
| Not always well structured and is semantically ambiguous. | Has a well-defined structure and semantics. |
| Does not provide a solution to the user of the database system. | Provides solutions to the user of the database system. |
| The results obtained are approximate matches. | The results obtained are exact matches. |
| Results are ordered by relevance. | Results are unordered by relevance. |
| It is a probabilistic model. | It is a deterministic model. |

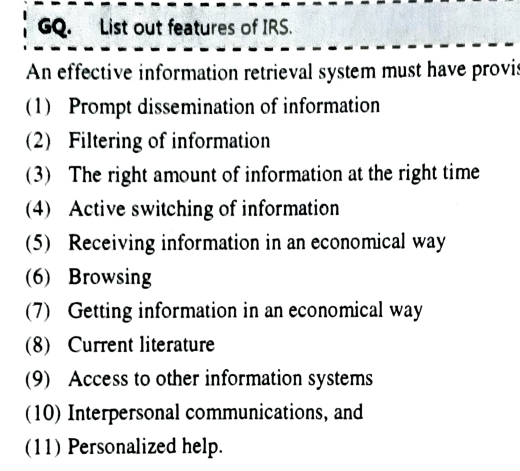
Objectives:

IRS is focused on two fundamental issues: Information storage and retrieval techniques.

* Retrieving the required information is a major goal of an information retrieval system.

It could be actual information or information surrogates from documents that completely or partially match the user's search.





What are the types of information retrieval?

Methods/Techniques in which information retrieval techniques are employed include:

1. Adversarial information retrieval
2. Automatic summarization
3. Multi-document summarization
4. Compound term processing
5. Cross-lingual retrieval
6. Document classification
7. Spam filtering
8. Question answering

Components of IRS

1. Inverted index
2. Stop word elimination
3. Stemming
4. Crawling
5. Query
6. Relevance feedback

A search engine is a kind of website through which users can search the content available on the Internet. For this purpose, users enter the desired keywords into the search field. Then the search engine looks through its index for relevant web pages and displays them in the form of a list. The Internet is a huge source of information & resources and to access the resource from the Internet there are some kinds of software, this software is known as a Search Engine. Some of the popular ones are Google, Bing, Yahoo, Duck duck go, Baidu, etc.

Components of search engine:

1. Crawler
2. Indexer
3. Search algorithm

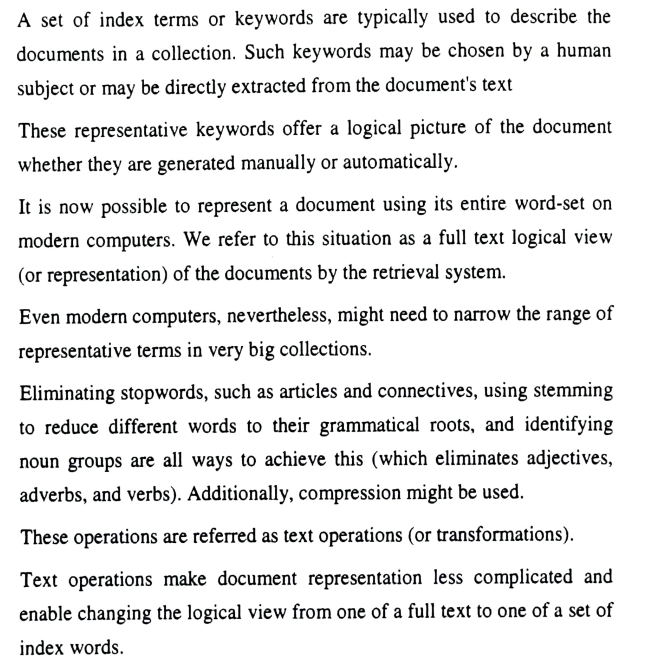
Web Browser:

The web browser is an example of application software that is developed to retrieve and view the information from web pages or HTML files present on the web servers. The first web browser was invented by Sir Tim Berners-Lee in 1990 and the very first graphical web browser was developed in 1993 and is named the mosaic. After that, various web browsers were developed. Some of them are navigator which is developed by Netscape communication, Microsoft Edge, Google Chrome, Mozilla Firefox, Opera, and Apple safari.

The main characteristics of a Web Browser are:

1. It consists of Graphical User Interface.
2. It contains the search box where the user can type the address or URL.
3. Page style can be static or dynamic. It depends upon the interactivity and the formatting.
4. TCP/IP and HTTP protocols are used by web browsers.

Logical View of the document

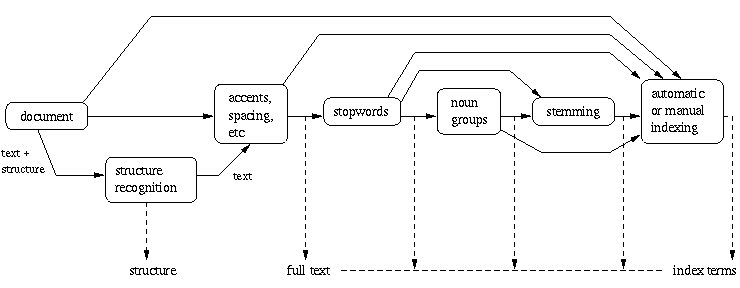


Due to historical reasons, documents in a collection are frequently represented through a set of index terms or keywords. Such keywords might be extracted directly from the text of the document or might be specified by a human subject (as frequently done in the information sciences arena). No matter whether these representative keywords are derived automatically or generated by a specialist, they provide a logical view of the documents.

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 (or representation) of the documents. With very large collections, however, even modern computers might have to reduce the set of representative keywords. 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). Further, compression might be employed. These operations are called text operations (or transformations) and are covered in detail in Chapter 7. 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 .

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 as illustrated in Figure . 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.). This information on the structure of the document might be quite useful and is required by structured text retrieval models such as those discussed in Chapter 2.

As illustrated in Figure , we view the issue of logically representing a document as a continuum in which the logical view of a document might shift (smoothly) from a full text representation to a higher level representation specified by a human subject.



Retrieval Process

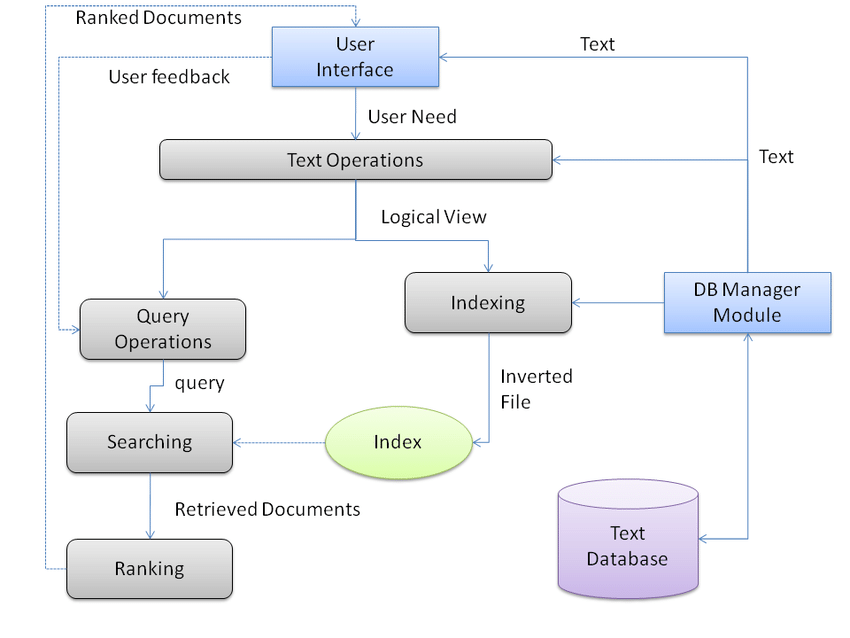
To describe the retrieval process, we use a simple and generic software architecture as shown in Figure . First of all, before the retrieval process can even be initiated, it is necessary to define the text database. This is usually done by the manager of the database, which specifies the following: (a) the documents to be used, (b) the operations to be performed on the text, and (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. An index is a critical data structure because it allows fast searching over large volumes of data. Different index structures might be used, but the most popular one is the inverted file as indicated in Figure . 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.

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

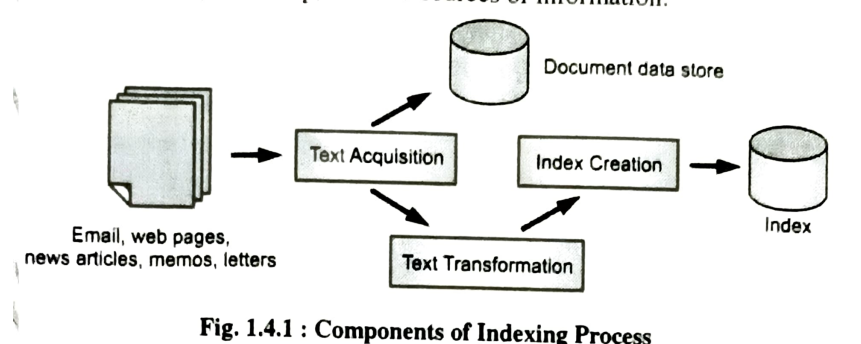
Given that the document database is indexed, the retrieval process can be initiated. The user first specifies a user need which is then parsed and transformed by the same text operations applied to the text. Then, query operations might be applied before the actual query, which provides a system representation 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.

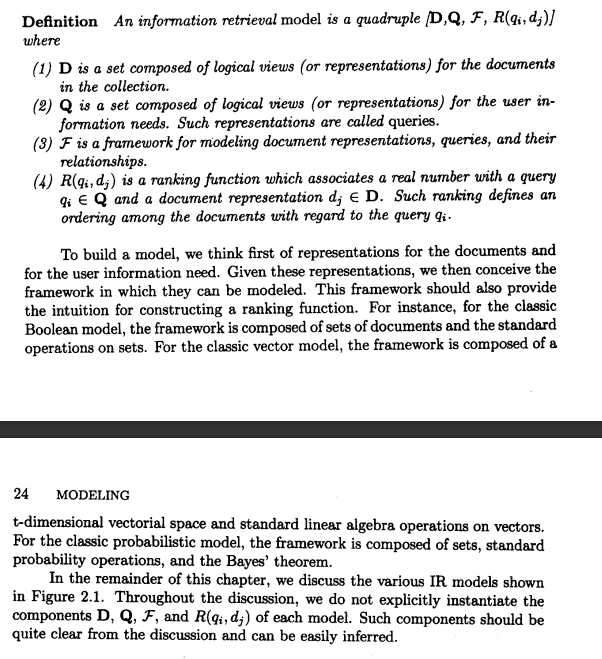
Before been sent to the user, the retrieved documents are ranked according to a likelihood of relevance. The user then examines the set of ranked documents in the search for useful information. At this point, he might pinpoint a subset of the documents seen as definitely of interest and initiate a user feedback cycle. In such a cycle, the system uses the documents selected by the user to change the query formulation. Hopefully, this modified query is a better representation of the real user need.



Types of IRS:

1. Operations Support System
2. TPS
3. MIS
4. DSS
5. Experts system
6. OAS



Formal characterization of IR Models  


Boolean - set theoretic - simple retrieval model based on set theory and Boolean algebra

Vector - algebraic - mathematical and geometric approach

Probablistic - probabilistic

Ad Hoc Retrieval: Ad hoc retrieval refers to the retrieval of documents from a collection based on user queries submitted at the moment (ad hoc) without any prior knowledge or filtering.

Web search engines like Google primarily use ad hoc retrieval. When you enter a search query, the search engine retrieves web pages that match your query, regardless of whether they were previously seen or indexed.

Filtering retrieval, also known as filtering-based information retrieval, focuses on the continuous monitoring and filtering of documents or information streams to deliver relevant content to users.

Email spam filters, news recommendation systems, and social media content feeds are examples of filtering retrieval. For instance, an email spam filter automatically filters out unwanted or irrelevant emails from your inbox based on predefined criteria.