RGPVNOTES.IN

Subject Name: **Modern Information Retrieva**

Subject Code: **CS-7004**

Semester: **7th**

UNIT – 1

Syllabus

Introduction: Information versus data retrieval, the retrieval process, taxonomy of Information Retrieval Models.

1.1 Motivation of Information Retrieval

Information retrieval (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 is interested. Unfortunately, characterization of the user information need is not a simple problem. Consider, for instance, the following hypothetical user information need in the context of the World Wide Web (or just the Web):

Find all the pages (documents) containing information on college tennis teams which:

1. Are maintained by a university in the USA.
2. Participate in the NCAA tennis tournament. To be relevant, the page must include information on the national ranking of the team in the last three years and the email or phone number of the team coach.

Clearly, this full description of the user information need cannot be used directly to request information using the current interfaces of Web search engines. Instead, the user must first translate this information need into a query which can be processed by the search engine (or IR system).

In its most common form, this translation yields a set of keywords (or index terms) which summarizes the description of the user information need. 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.

1.2 Information versus Data Retrieval

Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need. In fact, the user of an IR system is concerned more with retrieving information about a subject than with retrieving data which satisfies a given query. A data retrieval language aims at retrieving all objects which satisfy clearly defined conditions such as those in a regular expression or in a relational algebra expression. Thus, for a data retrieval system, a single erroneous object among a thousand retrieved objects means total failure. For an information retrieval system, however, the retrieved objects might be inaccurate and small errors are likely to go unnoticed. The main reason for this difference is that information retrieval usually deals with natural language text which is not always well structured and could be semantically ambiguous. On the other hand, a data retrieval system (such as a relational database) deals with data that has a well defined structure and semantics.

Data retrieval, while providing a solution to the user of a database system, does not solve the problem of retrieving information about a subject or topic. To be effective in its attempt to satisfy the user information need, the IR system must somehow `interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query. This `interpretation' of a document content involves extracting syntactic and semantic information from the document text and using this information to match the user information need. The difficulty is not only knowing how to extract this information but also knowing how to use it to decide relevance. Thus, the notion of relevance is at the center of information retrieval. In fact, the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible.

1.3 Basic Concepts

The effective retrieval of relevant information is directly affected both by the user task and by the logical view of the documents adopted by the retrieval system.

1.3.1 The User Task

The user of 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 the information need. With a data retrieval system, a query expression (such as, for instance, a regular expression) is used to convey the constraints that must be satisfied by objects in the answer set. In both cases, we say that the user searches for useful information executing a retrieval task.

Consider now a user who has an interest which is either poorly defined or which is inherently broad. For instance, the user might be interested in documents about car racing in general. In this situation, the user might use an interactive interface to simply look around in the collection for documents related to car racing. For instance, he might find interesting documents about Formula 1 racing, about car manufacturers, or about the `24 Hours of Le Mans.' Furthermore, while reading about the `24 Hours of Le Mans', he might turn his attention to a document which provides directions to Le Mans and, from there, to documents which cover tourism in France. In this situation, we say that the user is browsing the documents in the collection, not searching. It is still a process of retrieving information, but one whose main objectives are not clearly defined in the beginning and whose purpose might change during the interaction with the system.

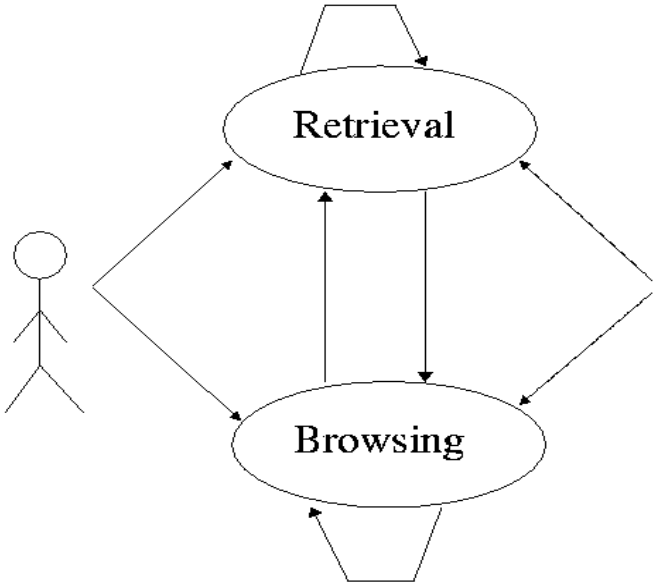

Figure-1 Interaction of the user with the retrieval system through distinct tasks.

Classic information retrieval systems normally allow information or data retrieval. Hypertext systems are usually tuned for providing quick browsing. Modern digital library and Web interfaces might attempt to combine these tasks to provide improved retrieval capabilities. However, combination of retrieval and browsing is not yet a well established approach and is not the dominant paradigm.

Figure 1 illustrates the interaction of the user through the different tasks we identify. Information and data retrieval are usually provided by most modern information retrieval systems (such as Web interfaces). Further, such systems might also provide some (still limited) form of browsing. While combining information and data retrieval with browsing is not yet a common practice, it might become so in the future.

Both retrieval and browsing are, in the language of the World Wide Web, `pulling' actions. That is, the user requests the information in an interactive manner. An alternative is to do retrieval in an automatic and permanent fashion using software agents which push the information towards the user. For instance, information useful to a user could be extracted periodically from a news service. In this case, we say that the IR system is executing a particular retrieval task which consists of filtering relevant information for later inspection by the user.

1.3.2 Logical View of the Documents

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 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 (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 stop words (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). 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 .

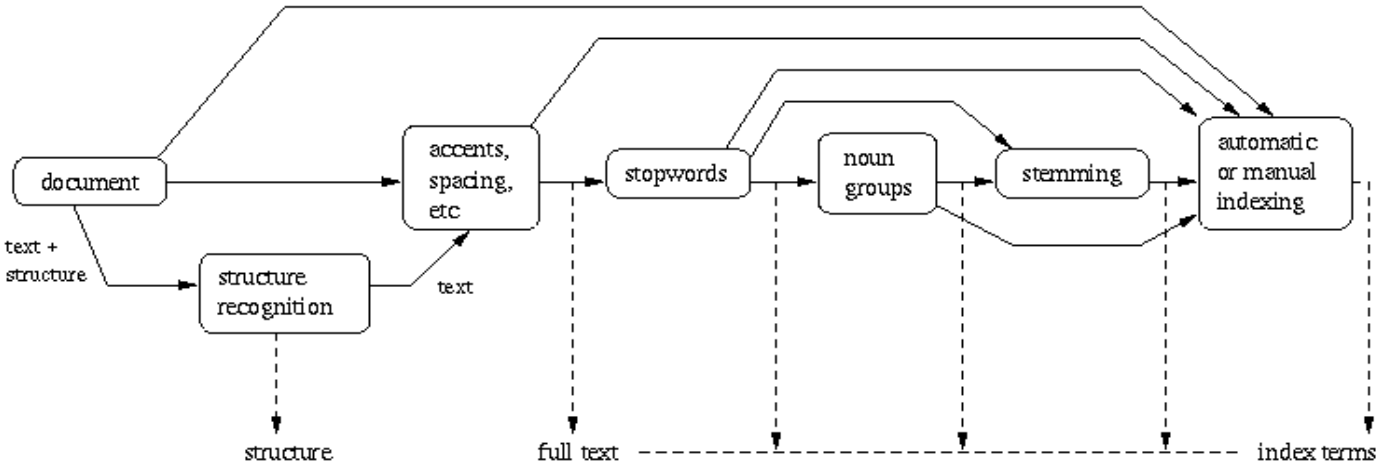

Figure 2: Logical view of a document: from full text to a set of index terms.

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 2**. 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.

As illustrated in **Figure 2**, 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.

1.4 The Retrieval Process

To describe the retrieval process, we use simple and generic software architecture as shown in **Figure 3**. 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 3**. 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.

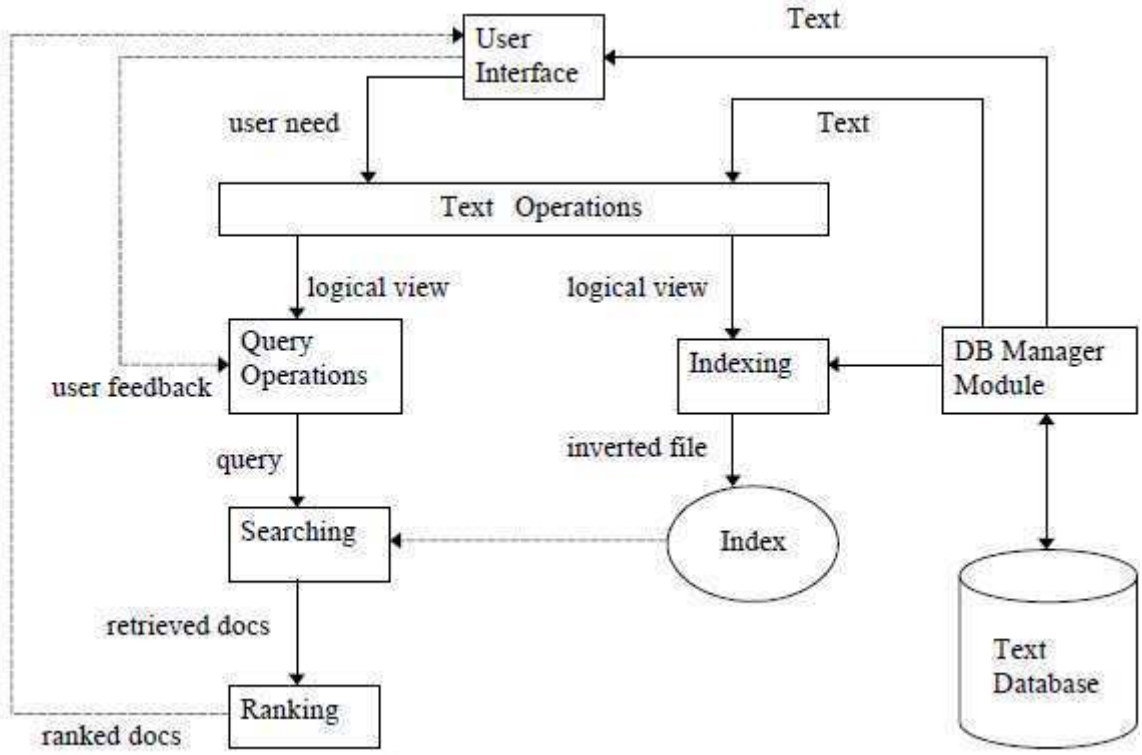

Figure 3: The Process of Retrieving Information

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.

1.5 Taxonomy of Information Retrieval Model

The 3 classic models in information retrieval are called Boolean, vector and probability. In the Boolean model documents and queries are represented as sets of index terms. So we can say that this model is set theoretic.

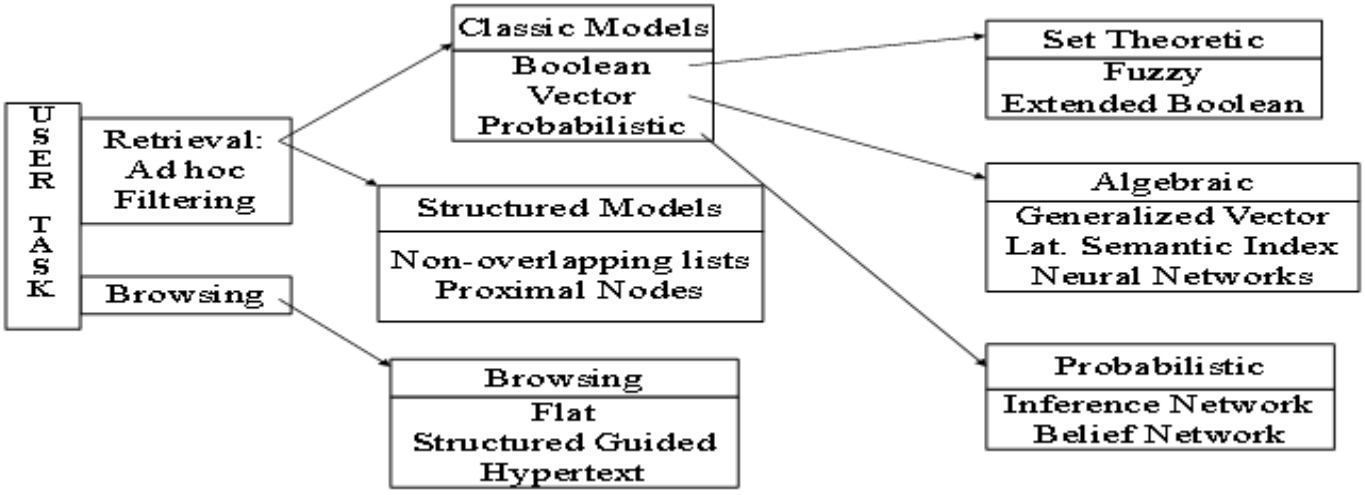

Figure 4: Taxonomy of Information Retrieval Model

In the vector model documents and queries are represented as vectors in t-dimensional space. Thus we can say that model is algebraic. In the probabilistic model, the framework for modelling document and query represe4ntation is based on probability theory. So we can say that the model is probabilistic.

Over the years alternative modelling paradigms for each type of classic model (i.e. set theoretic, algebraic and probabilistic) have been proposed. Regarding alternative set theoretic models, we distinguish the fuzzy and extended boolean models. Regarding alternative algebraic models, we distinguish the generalized vector, latent semantic indexing and the neural network models. Regarding alternative probabilistic models, we distinguish the inference network and belief network models.

Besides references to the text content, the model might also allow references to the structure normally present in written text. In this case, we say that we have structured model. We distinguish two models for structured text retrieval namely, the non overlapping lists model and proximal nodes model.

The user task might be one of browsing. We distinguish three models for browsing namely, flat, structure guided and hypertext.

	Index Terms	Full Text	Full Text+ Structure
Retrieval	Classic Set Theoretic Algebraic Probabilistic	Classic Set Theoretic Algebraic Probabilistic	Structured
Browsing	Flat	Flat Hypertext	Structure Guided Hypertext

Table 1: Retrieval models most frequently associated with distinct combinations of a document logical view and user task.

1.6 Retrieval: Ad hoc and Filtering

In conventional IR system, the documents in the collection remain relatively static while new queries are submitted to the system. This operational mode has been term ad hoc retrieval and is the most common form of user task.

A similar but different task is one in which the queries remain relatively static while new documents come into the system (and leave). This is the case with the stock market and with news wiring services. This operational mode has been termed filtering.

In a filtering task, a user profile describing the user's preferences is constructed. Such a profile is then compared to the incoming documents in an attempt to determine those which might be of interest to this particular user.

A variation of this procedure is to rank the filtered documents and show this ranking to the user. The motivation is that the user can examine a smaller number of documents if he assumes that the ones at the top of the ranking are more likely to be relevant. This variation of filtering is called routing but it is not popular.

RGPVNOTES.IN

We hope you find these notes useful.

You can get previous year question papers at https://qp.rgpvnotes.in .

If you have any queries or you want to submit your study notes please write us at rgpvnotes.in@gmail.com

LIKE & FOLLOW US ON FACEBOOK
facebook.com/rgpvnotes.in