UNIT V INFORMATION RETRIEVAL AND WEB SEARCH 9

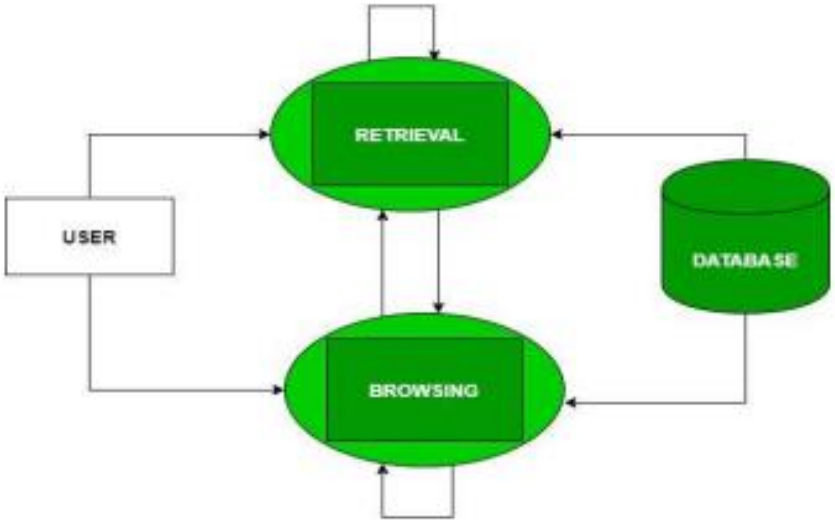
IR concepts – Retrieval Models – Queries in IR system – Text Preprocessing – Inverted Indexing – Evaluation Measures – Web Search and Analytics – Current trends.

IR CONCEPTS

Information Retrieval (IR) can be defined as a software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories, particularly textual information. 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.

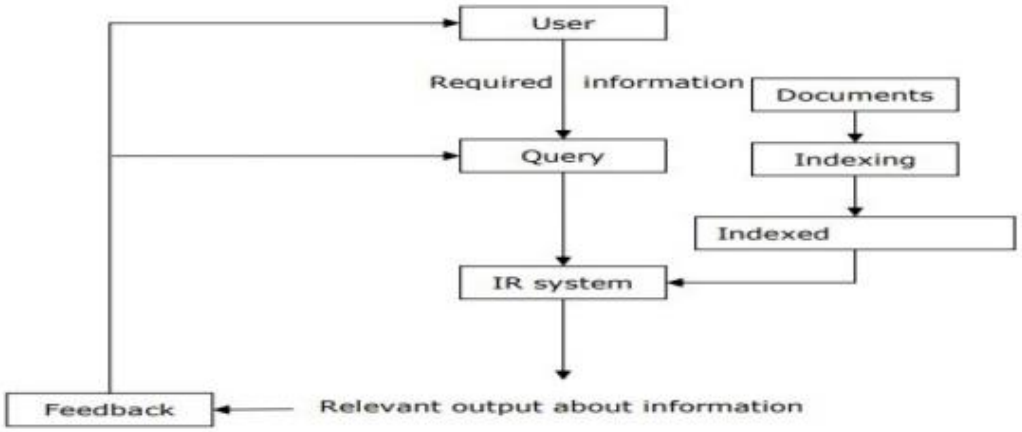
What is an IR Model?

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:



RETRIEVAL MODELS

It is the simplest and easiest to implement IR model. This model is based on mathematical knowledge that was easily recognized and understood as well. Boolean, Vector and Probabilistic are the three classical IR models. These are the three main statistical models—Boolean, vector space, and probabilistic—and the semantic model.



Types of retrieval model:

- Classical IR Model - It is the simplest and easy to implement IR model.
- Non-Classical IR Model - It is completely opposite to classical IR model
- Alternative IR Model
- Inverted Index.
- Stop Word Elimination.
- Stemming.
- Term Weighting.
- Term Frequency (tfij)

TYPES OF QUERIES IN IR SYSTEMS:

During the process of indexing, many keywords are associated with document set which contains words, phrases, date created, author names, and type of document. They are used by an IR system to build an inverted index which is then consulted during the search. The queries formulated by users are compared to the set of index keywords. Most IR systems also allow the use of Boolean and other operators to build a complex query. The query language with these operators enriches the expressiveness of a user's information needs.

1. Keyword Queries:

- Simplest and most common queries.
- The user enters just keyword combinations to retrieve documents.
- These keywords are connected by logical AND operator.

- All retrieval models provide support for keyword queries.

2. Boolean Queries:

- Some IR systems allow using +, -, AND, OR, NOT, ( ), Boolean operators in combination of keyword formulations.
- No ranking is involved because a document either satisfies such a query or does not satisfy it.
- A document is retrieved for Boolean query if it is logically true as exact match in document.

3. Phase Queries:

- When documents are represented using an inverted keyword index for searching, the relative order of items in the document is lost.
- To perform exact phase retrieval, these phases are encoded in an inverted index or implemented differently.
- This query consists of a sequence of words that make up a phase.It is generally enclosed within double quotes.

4. Proximity Queries:

- Proximity refers to search that accounts for how close within a record multiple items should be to each other.
- Most commonly used proximity search option is a phase search that requires terms to be in exact order.
- Other proximity operators can specify how close terms should be to each other. Some will specify the order of search terms.
- Search engines use various operators' names such as NEAR, ADJ (adjacent), or AFTER.
- However, providing support for complex proximity operators becomes expensive as it requires time-consuming pre-processing of documents and so it is suitable for smaller document collections rather than for web.

5. Wildcard Queries:

- It supports regular expressions and pattern matching-based searching in text. Retrieval models do not directly support this query type.
- In IR systems, certain kinds of wildcard search support may be implemented.
- Example: usually words ending with trailing characters.

6. Natural Language Queries:

- There are only a few natural language search engines that aim to understand the structure and meaning of queries written in natural language text, generally as questions or narratives.
- The system tries to formulate answers for these queries from retrieved results.
- Semantic models can provide support for this query type.
