EXPERIMENT 1

AIM: To understand the query structure and execute various structured queries.

THEORY:

There are a number of techniques to enhance the usefulness of the queries Some examples are the expansion of a word to the set of its synonyms or the use of a thesaurus Some words which are very frequent and do not carry meaning (called stopwords) may be removed We refer to words that can be used to match query terms as keywords.

Another issue is the subject of the retrieval unit the information retrieval system adopts The retrieval unit is the basic element which can be retrieved as an answer to a query We call the retrieval units simply documents, even if this reference can be used with different meanings.

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 need. The Information Retrieval (IR) system finds the relevant documents from a large data set according to the user query. Queries submitted by users to search engines might be ambiguous, concise and their meaning may change over time. Some of the types of Queries in IR systems are

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.

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.

Phrase Queries:

- When documents are represented using an inverted keyword index for searching, the relative order of items in document is lost.
- To perform exact phase retrieval, these phases are encoded in 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.

Proximity Queries:

- Proximity refers ti 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.

Wildcard Queries:

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

Natural Language Queries:

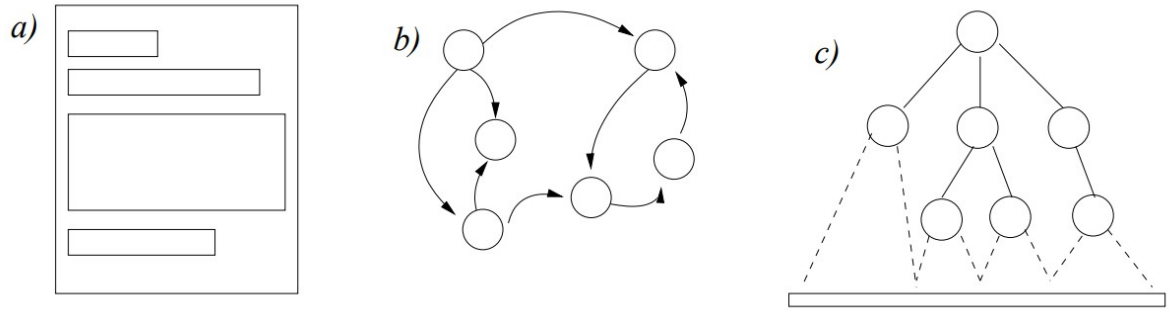
- There are only a few natural language search engines that aim to understand the structure and meaning of queries written in natural language text, generally as question or narrative.
- The system tries to formulate answers for these queries from retrieved results.
- Semantic models can provide support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

STRUCTURAL QUERIES:

- The text collections tend to have some structure built into them.
- The standardization of languages to represent structured texts has pushed forward in this direction.
- Mixing contents and structure in queries allows posing very powerful queries.
- Queries can be expressed using containment, proximity or other restrictions on the structural elements.

The three main types of structures:

- form-like fixed structure
- hypertext structure
- hierarchical structure



Fixed Structure:

The structure allowed in texts was traditionally quite restrictive. The documents had a fixed set of fields, and each field had some text inside

- Some fields were not present in all documents
- Some documents could have text not classified under any field
- They were not allowed to nest or overlap

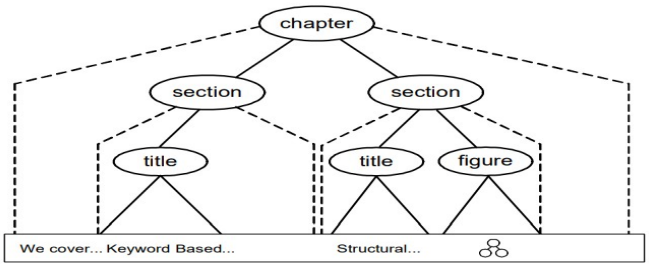
Retrieval activity allowed: specifying that a given basic pattern was to be found only in a given field. When the structure is very rigid, the content of some fields can be interpreted as numbers, dates, etc. This idea leads naturally to the relational model, each field corresponding to a column in the database table. There are several proposals that extend SQL to allow full-text retrieval. Among them we can mention proposals by the leading relational database vendors such as Oracle and Sybase, as well as SFQL.

Hypertext structure:

Hypertexts probably represent the opposite trend with respect to structuring power. Retrieval from hypertext began as a merely navigational activity. That is, the user had to manually traverse the hypertext nodes following links to search what he/she wanted. Some query tools allow querying hypertext based on their content and their structure.

Hierarchical structure:

An intermediate model which lies between fixed structure and hypertext is the hierarchical structure An example of a hierarchical structure: the page of a book and its schematic view



CONCLUSION: Hence we successfully understood the query structure and executed various structured queries.