

Information Retrieval
ITIS401
Chapter-7
Part-1
2021-2022
Dr Mohamed Abdeldaiem

Querying the Information Retrieval System

- Introduction
- Language Types
- Query Logic
 - 1- Sets and Subsets
 - 2- Relational Statements
 - 3- Boolean Query Logic
 - 4- Ranked and Fuzzy Sets
 - 5- Similarity Measures

Introduction

- If the query is in natural language, then the IRS interpreter must exercise a great deal of “judgment” in order to be able to translate accurately into the target command language.
- If the query is written in a typical procedural query language often known as a command language, little or no judgment may be required for the IRS to do faithfully what was asked.

Language Types

A query language is the means by which the user tells the IRS what to do or what is wanted. There are two broad types: procedural, and non-procedural or descriptive.

A procedural language involves the use of commands. It is much like a traditional computer programming language. The user makes such imperative statements as `SELECT` \updownarrow records meeting certain criteria \rightarrow , `PRINT` \updownarrow certain records from a designated set of the database \rightarrow , or `LOGOFF` from the current online session and break the communication link.

A non-procedural language is used to tell the IRS what result is wanted, and the IRS then works out how to produce it.

Figure.1

- Shows an example of a graphic query representation or concept map.
- The query represented is for a search on earthquake-resistant construction techniques used in California, with a date restriction.

Name of facet:	EARTHQUAKES	CONSTRUCTION	CALIFORNIA	1990-1999
Content of facet:	earthquakes tremor shock	construction building	california san andreas	1990-1999

Query Logic

- In most retrieval systems, users are asked to consider the database as a set of records about a set of entities.
- The retrieval procedure is to describe characteristics of subsets of the database in which the user is interested.
- Indeed, one definition of IR is that it is a method of partitioning a database such that one partition holds only the records of interest.

Sets and Subsets

- The design of most text-oriented retrieval systems, until very recently, anticipated frequent revision of the query statements and called for the IRS to create a set for each query, or even for each attribute-value specification or Boolean combination within the query.

Cont

- Each such set represented a subset of the database, but the word set is the common usage. Its physical manifestation is not seen by the user.
- It is a list of the identifying accession numbers of the retrieved records satisfying the query or component statement. The user is typically informed only of the set number assigned by the IRS and the number of records in the set. A set with no members is called a null set.

Cont

- Most Web search engines promptly display the found surrogates for any set formed and do not retain sets once displayed. Hence, to modify a set requires a modification of the query statement and a new search.

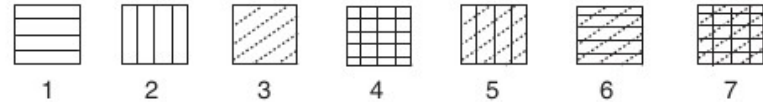
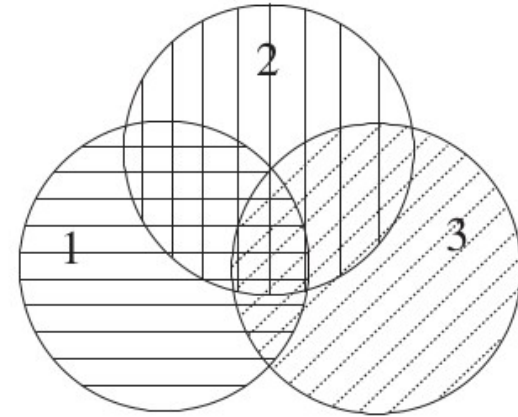
Relational Statements

- Relational statements specify the characteristics of records in a set to be formed or, equivalently, the characteristics of records that compose a subset of the database. Sets are defined by specifying one or more combinations of an attribute, a relationship, and a value; for example publication date \square 1987, author \square KRAFT , DONALD H . or salary $>$ 45,000.

Equality (=) (subject = TENNIS).
Inequality ($>$, $<$, \leq , $<>$, \geq) (date $>$ 881014 or subject $<>$ TENNIS). The symbol $<>$ is used for *not equal* in most computer languages.

Boolean Query Logic

- Historically the first and still the most common method used for expressing micro logic in query statements is Boolean algebra.
- This involves specifying the operations to be performed on sets that have been defined by relationship statements or previous set operations.
- The sets are all subsets of a universe of dis-course, which in this case is the database being searched.



cont

- Normally, experienced searchers do not expect to get the exact results they want on the first try, especially when searching records on the basis of words contained in a natural-language text.
- Therefore, their approach is to use a series of search statements or queries, beginning as probes and gradually improving on earlier results.

For example

- Suppose we are looking for information about the nutritional value of prepared foods for pet cats and dogs.
- It is worth a diversion here to point out that use of and in the preceding sentence has the meaning of or in Boolean logic.
- “Cats and dogs” in colloquial usage means, “either cats or dogs or both,” but a rigidly literal interpretation might be “what is in common to, or a member of both sets of, cats and dogs.”
- This is one of the reasons it can be difficult to teach Boolean logic applied to natural language. Now, the search might start with:

cont

SELECT PET OR PETS

(Creating Set 1, say of 52,000 records)

Next, try

SELECT DOGS OR CATS

(Creating Set 2 of 8000 records,
seemingly better than the 52,000 of
the first try)

SELECT FOOD AND
NUTRITION

(Set 3, 5000 records)

SELECT S2 AND S3

(Set 4, 3 records)

But this last may seem too small, so the searcher goes back to the more general

SELECT S1 AND S3

(Set 5, 27 records)

Ranked and Fuzzy Sets

- Ranking means to assign to each record a measure of the closeness of the record's content to the query, or the extent to which the record matches the query.
- If ranking logic is used, Boolean operations can still be used to define a set, but we shall not be satisfied to say merely that any record is either in the set or not.

cont

- The point of ranking is to acknowledge that there is uncertainty (the definition of the set is fuzzy) as to whether the query exactly expressed the user's needs.
- If the user knows the question is imperfect, there is little to be gained in getting back an answer that claims, in effect, "Here is the exact information you wanted."
- Mathematically, we can define a binary set membership function, S , such that for any record D_i and query Q , the functional value is either 1 or 0, depending on whether D_i satisfies Q or not.

cont

- A record satisfies a query if the attribute values and combinations of values requested in the query are found in the record.
- A notation can be
- $S(D_i \times Q) \rightarrow \{0,1\}$,

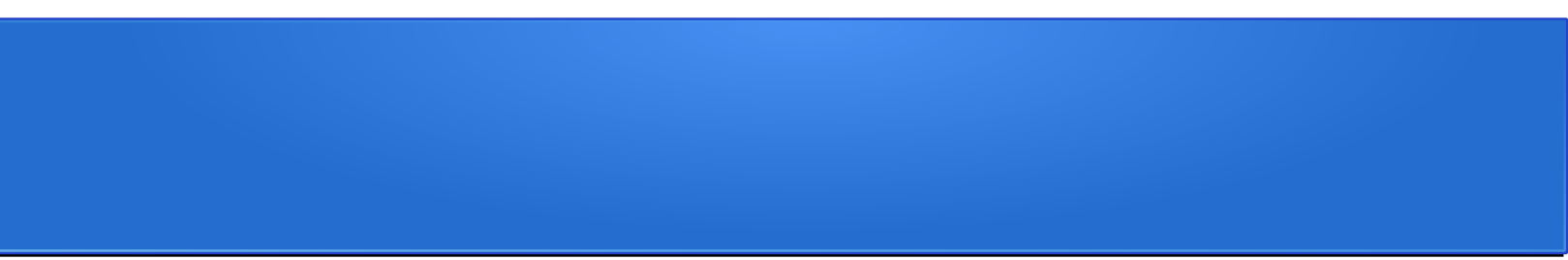
which says that S maps the combination of a record and a query into a set of values, each either 0 or 1. The curly brackets indicate a set whose members are selected from the values shown within.

cont

- Then the retrieval set for Q is all the records D_i such that $S(D_i \times Q) = 1$. This is a crisply defined set, each record of the database being clearly in the set or not in it.

Similarity Measures

- Another variation is relevance feedback (Salton and McGill, 1983, pp. 140–145, 236–243), the user provides a query statement, then rates some of the retrieved records on relevance.
- Using this rating, the IRS can increase or decrease the weighting of terms, select new ones, or delete unproductive ones in the original query on the basis of frequency of occurrence of terms in records.



Information Retrieval
ITIS401
Chapter-7
Part-2
2021-2022
Dr Mohamed Abdeldaiem

Functions Performed (practical part)

- 1 Connect to an IRS
- 2 Select a Database
- 3 Search the Inverted File or Thesaurus
- 4 Create a Subset of the Database
- 5 Search for Strings
- 6 Analyze a Set
- 7 Sort, Display, and Format Records
- 8 Handle the Unstructured Record
- 9 Download
- 10 Order Documents
- 11 Save, Recall, and Edit Searches
- 12 Current Awareness Search
- 13 Cost Summary



The End of Chapter-7

Any Questions.....?