Querying of the Information Retrieval System

I. Objectives:

The main objective of this module is to:

- Introduce the basic concepts of query language and its importance in information retrieval system.
- Introduce the concept of set and subset for formulation of query statement in IR system.
- Introduce to the reader to relational operators, which are used for formulation of query.
- Introduce reader to the different types of Boolean operators and their function in IR.
- Introduce the students to basic concept of fuzzy information retrieval for fuzzy document evaluations and fuzzy queries.
- Describe the concept of similarity measure in information storage and retrieval.

II. Learning Outcomes

After reading this module:

- Students will gain knowledge of Information Retrieval System and its role for formulating query in search and retrieve process .
- Students will understand the meaning, types and function of query languages.
- The reader will able to differentiate between command language and natural language/non-procedural language.
- The learner will gain knowledge of Boolean operators in forming query sets in IR system.
- The reader will also gain knowledge of fuzzy retrieval and similarity measures in document retrieval and for query formation.

III. Structure

1. Introduction
2. Sets and Subsets
3. Relational Statements
4. Boolean Query Logic
5. Ranked and Fuzzy Sets
6. Similarity Measures
7. Summary
8. References

1. Introduction

Information search and retrieval involves finding and retrieving useful documents from a store of information. In any information search and retrieval system an important factor which plays a role is search and selection process. Users issue a query to Information Retrieval System (IRS) to find useful documents answering their information need, from a large volume of information.

Information search can be made by presenting a query through the inter-mediator or directly to the IRS. A query in general terms is a statement or series of statements made by a user to a retrieval system for the purpose of specifying what information is to be retrieved and in what form.

Most of the time the query is specified in a format using a language called 'query language'. A query language is the means by which the user tells the IRS what to do and what is wanted. A query is distinct from the types of documents that the user is trying to retrieve. The document representation and the query undergo parallel processes within the retrieval system. On the document side, someone generates or gathers some data and formulates it into a document representation for example an annotation, metadata etc. After creating document representations, they are transferred into internal representation, which then gets transferred into a format that is used for matching process. On the query side, user begins with information needs.

There are two broad types of query language: procedural and non-procedural or descriptive. A procedural language uses commands. If the query is written in a typical procedural query language often known as command language, little or no knowledge is required for the IRS to find what was asked for and retrieve the same. Natural language queries or non-procedural queries generally tend to be ambiguous in syntax and meaning. For natural language queries inter-mediator is required to formulate a query as these queries generally tend to be ambiguous. Command language queries are more structured and for IRS these queries are unambiguous when compared to natural language queries.

1.1 Querying Distinctions between IR Systems and Database Management Systems

One of the most common questions is how an IR system is different from a Database Management system from querying and retrieval point of view. It may seem that database management system is quite similar to IR system but there are significant differences between them. The difference generally is there because of the purpose they were used for and according to the stated purpose they perform the different functions.

Database management systems generally store, retrieve and present the particular data facts such as name, date of birth and thus contain similar design and query features but IR system are much more loose in nature and hence handle variable length entries easily. Repetition of field entries is handled more subtly in IR systems. IR systems are capable of handling less structured content especially for full text documents. DBMS have a lot of data models which helps to store data securely and with integrity where as in IR systems since data is stored flat-file databases it causes data redundancy and integrity issues also. The most important difference is the way the search results are evaluated. Though both of them allow querying they are at distinct levels of detail.

DBMS generally index down at field level, thus searches of words if required within the field becomes difficult where as IR systems does index down to the word level within the field. They include advanced search features proximity searching and relevance ranking also. The concept of relevance is quite unique to IR systems and while general DBMS may not include this. **(Wolfram, 2003)**

The objective of this module is to introduce the basic concepts of querying a information retrieval system. So , in the section 2, Set and Subsets are described which are helpful in creating a small store of documents which retrieves back the results when subjected with same query time and again. In the section 3, Relational Statements are described which are made with the help of relational operators to formulate a query. Section 4 , describes about the boolean query logic which provides various ways of querying and helpful for creating different kind of Set and Subsets. Section 5 , describes about Fuzzy set theory and basic operations, which helps in overcoming some of the limitations of boolean query logic. Section 6,describes about Similarity measures basics and its different kinds. In the final section summary of this module is given.

2. Sets and Subsets

Most text information retrieval systems are designed in such a way that it is anticipated that user will make frequent revisions in the formulation of query statement and hence the IRS will create and maintain a set and subset of each query which generally include Boolean combinations within the query. Set is also generally a list of identifying accession numbers of the retrieved records satisfying the query or component statement. The user is informed only of the set number assigned by the IRS and the number of records found in the set. A set with no number is generally termed as a null set. To modify the search one has to modify a query statement and carry out new search.

3 Relational Statements

Relational statements specify the characteristics of records in a set to be formed or 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 = 2000, author= Croft, B. or salary > 10,000. The relational statements in any IRS include

Equality (=) e. g. subject=library science. Equality is not only the relationship that is expressed. There is an inequality characteristic as well in any relational statement. Inequality is expressed by the following symbols - (>,<,<=,>=, <>) e. g. (date > 10102012 or subject <> library science). The symbol <> is used for not equal in most computer languages.

3.1 Relational operators

The symbols =,>,<,<=,=>,<> are called as relational operators which establish relation between two entities. Each of the symbols has two operands to work with which could be string or numbers. All relational operators have equal preference but they are lower than those of arithmetic operators. They provides us with additional options for querying the system. These relation operators are very important in selecting the Sets and Subsets from a system and the

operands play crucial role. The operators are mainly used in programming languages to write codes so that proper data or information can be retrieved from a system. The relational operators are also used to test if a particular element is present in the retrieved set or not. Thus they help in clustering similar kind of information. **(Larsen, K. S., Schmidt, E. M., & Schwartzbach, M. I. , 1989)**

4 Boolean Query Logic

A user often thinks of a query in one of the two forms – either as a question in everyday language [English or some other natural language], or as a list of terms. A query developed from a list of terms is called a Boolean Algebra. Boolean algebra is historically first and still the most common method used for expressing a query statement. Boolean query involves specifying the operation to be performed on the query sets that have been defined by relationship statements as mentioned above.

Boolean query is also based on concepts from logic with its terms joined together by logical connectives. Typically the connectives permitted are AND, OR, and NOT. Together with grouping of terms, these operators are sufficient to express any logical combination of terms.

A typical Boolean AND logic query might be like 'Library AND Information'. Boolean operators can be combined in different ways to express information need in a query. One more Boolean operators can be used in a query to specify requirements. Eg., Digital AND (Library OR Information) AND Management. The appropriate response to this query is to retrieve any document having the words Digital Library Information Management.

In a Boolean query the use of AND requires that both the terms be present in the retrieved documents. In the above diagram given sets 1 and 2, a third set can be defined that includes only records that are in both sets 1 and 2. If set 1 contains records with say date > 10102012 and set 2 contains records with subject = library then set 3 can be defined as SET 1 and SET 2, which would contain records with both the desired key terms such as date range and the desired subject. This is called the logical product, intersection, or conjunction of the first two sets.

The use of OR operator requires that at least one of the term be present in the search results. Given two sets a third can be formed from records that are in either the first or the second or both. This is written as SET 1 OR SET 2 and is called logical sum, union or disjunction of the first two sets. OR operator is always used in the inclusive sense, meaning records that are in the set 1 or set 2 or both are retrieved.

The use of NOT operator requires that the specified term be absent from any retrieved document. While AND and OR are binary operators in the sense that their operations require at least two sets within the universe of discourse, NOT is a unary operator, which means it can be applied on a single set. For e.g. If the single set is A, then NOT A is known as the complement of A. Normally in a retrieval context, NOT is considered to mean AND NOT, and is treated as a binary operator. For instance, if one needs records of 'library science' but not about 'social science', the logic is that set 1 is to contain records with library science ; set 2, records without social science ; and set 3, their intersection, records containing library science but not social science.

Symbolically, SET 3 = SET 1 AND NOT SET 2. Set 3 is called the logical difference of sets 1 and 2.

4.1 Boolean Search Modifier:

There are some modifiers that are deployed to narrow down or expand the scope of queries. Some example modifiers are:

a) ASTERISK *

The retrieval system when queried using the * asterisk would give back all the words that starts with the root/stem word which has been cut short by the asterisk itself. **(monyl, 2008)**

For eg: Freeze* will return: freezer, freezing, etc.

This helps the user as the user may know only part of a term or he/she is not sure of the entire spelling. In such cases the feature is useful but the user has to ready for some noise [irrelevant resources] in retrieval also.

b) PARENTHESES

It is considered as a best practice to use parentheses while writing OR statements, because it helps to execute OR statements properly. **(monyl, 2008)**

Eg., (A OR B) AND C

c) QUOTATION MARKS ""

"" are used for exact phrase search. When users are typing multi-worded query and some of the words are part of a single term, it is a best practice to use the quotation marks otherwise the IR systems might split the phrase and treat them like lone words and thus results will be less precise

For eg., "Information Retrieval System" retrieves resources where these three terms are exactly like that. In the absence of the double quotes the three words are treated as separate words and recall will be more; precision will be less.**(monyl, 2008)**

4.2 Limitations of Boolean query

Despite the simplicity and appeal of Boolean queries, it presents a number of significant problems. In a pure Boolean query there is no good way to weight terms for significance. Either a term is present or absent. Thus the user has little control over how important a given term is to the query. The second problem with Boolean queries is retrieval results can be wrong because of a misstated query. The third problem with Boolean query is that the user is free to construct a highly complex query that requires the development of many partial responses to be assembled

into the final response to the query. The fourth problem associated with Boolean retrieval systems is in controlling the size and composition of the retrieved set. Technically, the system should return all documents satisfying the query. This may, however, be either a very small number or a very large one. While executing a Boolean query, sorting the documents by the number of matching query terms provides a rough ranking of the documents. Boolean retrieval systems are highly popular and reasonably effective because of its ease of use in IRS.

5 Ranked and Fuzzy Sets

Ranking means to assign 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 operators can still be used to define a set. The purpose of ranking records is to acknowledge that there is uncertainty as to whether the query exactly expressed user needs or not.

In case of fuzzy information retrieval when a user cannot accurately tell whether a given document will meet the information need this uncertainty is called of “fuzzy” evaluation of the document with respect to the query. The concept of fuzzy information retrieval allows for both fuzzy document evaluations and fuzzy queries.

5.1 Fuzzy Logic

Fuzzy logic, rather than being based on Boolean (0 or 1, true or false) logic, it is the process of finding the correctness based on modern computer approaches. In information retrieval, this technique is basically used to solve the problem of natural language processing. For example, In Boolean logic it is not always possible to convert any problem into 0 or 1 form. Sometimes some problem will arise between 0 and 1. So to solve those kind of problem we need fuzzy logic technique. This technique was first proposed by Dr. Lotfi Zadeh of the University of California, Berkeley in the 1960s.**(Fuzzy logic, n.d.)**

5.2 Fuzzy Set

A Set is a collection of things. A fuzzy set is a set of things which may not belong to either one category. For instance set of 'old men'. Here we may define those who are aged greater than or equal to 60 years are old. So we try to convert this into 0 or 1, but there are many cases that cannot be included, like suppose there are set of people whose are 60.1 year old and some are 59.9. So in this case it is difficult tell who is old and who is not old. The fuzzy set technique provides us a better way to solve this kind of problem. **(“Fuzzy sets”,n.d.)**

5.3 Fuzzy Set Operations

Similar tp Boolean Operations (AND, OR, NOT), fuzzy set also have same operations not only for finite number but also for infinite large numbers. It means fuzzy set not only include finite value 0 or 1 but also the values between (0, 1). **(“Fuzzy logic: The logic of fuzzy sets,” n.d.)**

Consider the two sets A and B. The operators can be expressed as below:

AND Operation

The AND operation in fuzzy logic is min (A, B)

A	B	A AND B
0	0	0
0	1	0
1	0	0
1	1	1
0.1	0.5	0.1
0.2	0.7	0.2
0.3	0.8	0.3

OR Operation

The OR operation in fuzzy logic max (A, B)

A	B	A OR B
0	0	0
0	1	1
1	0	1
1	1	1
0.1	0.5	0.5
0.2	0.7	0.7
0.3	0.8	0.8

NOT Operation

This is means in fuzzy operation is 1-A

A	NOT A
0	1
1	0
0.2	0.8
0.3	0.7
0.5	0.5
0.9	0.1
0.8	0.2

6 Similarity Measures

Similarity is another important concept in information storage and retrieval. The aim of IRS is to retrieve those documents whose contents are similar to the information need as mentioned in the query formed. To aid in this process, cataloguers and indexers try to organize the document collection so that similar documents are in some sense close together and can be retrieved as a group with minimum effort and time.

A document can be represented by a list of terms that occur in it. A common way to define document similarity is to relate it to the key terms that two documents have. In order to handle the documents within the collection, it is assumed that all the terms in collection are in fixed order say 'alphabetical order'.

In case of retrieval of records, number of co-occurring words serve as a measure of similarity between two texts and the percentage of co-occurring words better serves the purpose of searching. Precision of retrieval can be achieved by considering the number of words in common and the number not in common. There are several ways to measure how similar two texts are. They all use the number of terms in common to the two texts, but other factors also play an important role such as sizes of the documents involved, the number of terms not in common and weights that may be assigned to the terms. The use of weights, allows greater importance to be given to co-occurrence of highly weighted terms. The use of weights however can be highly subjective. A similarity measure is generally applied to pairs of documents or to a document and a query.

6.1 Distance Based Similarity Measure

It is one of the most popular similarity measure. The idea is that the mental representation of a object, concept termed as 'percept' can be quantified with various features. Therefore if we have a imaginary space and represent each of the percepts with numerical values then these values can be considered as the coordinates of those percepts. **(Ashby & Ennis, 2007)**

Since a percept may have a lot of features we can use the Multidimensional scaling (MDS) . It is a method of performing similarity measure where, in an imaginary space similarity of percepts is inversely proportional to the distance between the percepts. Examples of distance based similarity measures in MDS are 'Euclidean distance' and 'City-block' distance.

6.2 Feature Based Similarity Measure

This similarity measure argues against the distance similarity. It is based on the feature matching process which tries to find similar and distinct features of the percepts in question.

If $a(x \cap y)$ denote the features which are common to both x and y and let $a(x - y)$,$a(y-x)$ denote the features which are only for x and y respectively, **Tversky's** (1977) feature contrast model proposes that the similarity of x to y is equal to

$$r(x, y) = \alpha\, a(x \cap y) - \beta\, a(x - y) - \gamma\, a(x - y),$$

where α, β, and γ are constants that might vary across individuals, context, and instructions. This model states that increase in number of common features increases the similarity and increase in number of unique features of each x and y decreases the similarity measure. This model helps us to find the deviations in the distance axioms.

6.3 Probabilistic Similarity Measure

The above mentioned similarity measures i.e in section 6.1,6.2 consider that percepts can be determined. But many researchers say that because the features of percept change over time, it is very difficult to say how concrete they can be, thus the concept of probability comes. There have many probabilistic models but all of them generally have two assumptions in common which was given by L. L. Thurstone (1927) and by signal detection theory (Tanner & Swets, 1954) that

i. If we expose the feature time and again to the percept in question, the percepts may draw out different results

ii. "there is a well-defined decision rule that describes how a response is selected for any momentary value of the percept". Here it is important to note that it is not that researchers discard the distance similarity measure in the probabilistic models; some of the do accept that similarity is inversely related to the distance measured.

7. Summary

An Information Retrieval System is only as good as its capacity to retrieve precisely relevant resources in reply to a query. But again the query itself should reflect the actual information need of users precisely. In this module types and function of query languages is described. The differences between command language and natural language/non-procedural language are explained. Boolean logic and its role in formulating query in search and retrieve process are covered. Various Boolean operators in forming query sets in IR system are illustrated. Advanced topics in query formulation such as fuzzy retrieval and similarity measures in document retrieval are also explained.

8. References

1. Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. New York: ACM Press.

2. Fuzzy logic: The logic of fuzzy sets. Retrieved October 5, 2016, from http://www.sjsu.edu/faculty/watkins/fuzzysets.htm

3. Fuzzy Sets. (n.d.). Retrieved October 8, 2016, from https://www.calvin.edu/~pribeiro/othrlnks/Fuzzy/fuzzysets.htm

4. Information storage and retrieval by Korfhage, R. R., Wiley Computer Publishing, New York, 1997, 349 p. ISBN: 0-471-14338-3

5. Karn, B. INFORMATION RETRIEVAL SYSTEM USING FUZZY SET THEORY - THE BASIC CONCEPT. Retrieved from http://pchats.tripod.com/istebhaskar.pdf

6. Kirankumar, B., Prasad, D. S., Manohar, M., Satyaprakash, K., Chiranjeevi, M., & Kiran, V. K. (2012). DATABASE MANAGEMENT SYSTEM AND INFORMATION RETRIVAL. Retrieved from http://www.ijcsit.com/docs/Volume%203/Vol3Issue2/ijcsit2012030270.pdf

7. Larsen, K. S., Schmidt, E. M., & Schwartzbach, M. I. (1989). A universal relational operator. Aarhus, Denmark: Aarhus University, Computer Science Dept.

8. Lin, D. (1998, July). An information-theoretic definition of similarity. In ICML (Vol. 98, pp. 296-304).

9. monyl. (2008, December 19). Basic Boolean search operators and query modifiers explained. Retrieved October 10, 2016, from Boolean, http://booleanblackbelt.com/2008/12/basic-boolean-search-operators-and-query-modifiers-explained/

10. Nowacka, K., Zadro»ny, S., Newelska, U., & Kacprzyk, J. A new fuzzy logic based information retrieval model *. Retrieved from http://www.gimac.uma.es/ipmu08/proceedings/papers/234-Zadrozni.pdf

11. Similarity measures. (n.d.). Retrieved October 10, 2016, from http://www.scholarpedia.org/article/Similarity_measures

12. Text information retrieval systems by Meadow, C. T., Boyce, B. R., U. K., Emerald, 2007, 371

13. What is fuzzy logic? - Definition from WhatIs.com. (n.d.). Retrieved October 16, 2016, from http://whatis.techtarget.com/definition/fuzzy-logic