A NEED TO EXPLORE THE POSSIBILITIES OF A CONTROLLED VOCABULARY/ FREE-TEXT INDEXING, IN THE SYSTEM WHICH FACILITATES THE INFORMATION SEARCHING ACTIVITY

D.C.Kuruppu (BSc., MSc.,M.Phil.)
Senior Asst. Librarian,
Faculty of Science, University of Colombo,
Colombo, Sri Lanka.

Abstract

Information retrieval is the process of searching for information and data, stored in or contained in various formats and sources. Any system, which facilitates the information searching activity, is normally called an information retrieval system. The system is made of six major subsystems, which are the document selection subsystem, the indexing subsystem, the vocabulary subsystem, and the subsystem of user-system interaction and the matching subsystem. The indexing subsystem consists of a bibliographic description of each document in the collection along with various access points to make this retrievable. Controlled vocabulary is considered as an important index variable and also plays a central role in a typical retrieval system. This means that vocabulary control is a relative rather than an absolute system. Modern information retrieval systems generally make use of free-text indexing (e.g.; database on On-line system). These two indexing systems have some artificial barrier between users and information. This paper briefly discusses the advantages and the disadvantages of controlled vocabulary and free-text indexing.

Introduction

An efficient, modern information service should be able to guarantee that any data contained in any document should be accessible to any member of the user community served. The ability of an information centre to supply known items is its "document delivery capability". The ability of the centre to retrieve documents on a particular subject, or provide the answer to a specific question is its "information retrieval capability." These two functions, document delivery and information retrieval are the major activities engaged in information services. Information retrieval is the process of searching some collections of documents, databases and other sources in order to identify a particular subject by using the terms in their widest sense.

When the documents are acquired by the Information Centre, they should be organized and controlled so that they can be identified and located. Activities include classification, cataloguing, subject indexing and abstracting. The subject indexing process involves two intellectual steps: "conceptual analysis" and the "translation" of the conceptual analysis. For efficient conceptual analysis, the indexer needs an understanding of the document, that is, some comprehension of its subject matter and a good knowledge of the needs of the users of the system. The recognition of what aspects of the document are most concerned can provide the constituents of conceptual analysis. The second step in the indexing process is the translation of the conceptual analysis into some vocabulary or "index language".

There is a limited set of terms that must be used to represent subject matter of documents such as the vocabulary which might be used to represent subject headings, a classification scheme, a thesaurus, or simply a list of approved key words or phrases. An "uncontrolled vocabulary" places no restrictions on the terms the indexer may use. In free-text indexing, all the terms will be included in the database index. This indexing is carried out by the computer. Whenever new records are added to the database, the computer updates the index to that database.

This paper briefly discusses the controlled vocabulary and the free-text indexing with their advantages and disadvantages.

Advantages and Disadvantages of a Controlled Vocabulary Vs Free-text indexing

The vocabulary of a retrieval system exerts a considerable influence on the information retrieval process. It greatly influences the construction of each search strategy and also greatly influences the attempt to match the actual search with the database. As a result of the lack of vocabulary control, the particular topic may be represented in many different ways in different documents or by different indexers. For example, the same pathological condition may be represented by one indexer as pulmonary tuberculosis and by another as tuberculosis of the lung. In an uncontrolled vocabulary situation the searcher must think of all synonymous words or expressions in order to find all relevant literature on a particular topic. Most information systems operate by means of a controlled vocabulary. A controlled vocabulary is essentially a set of terms that must be used at both the input and output stages of the information system. The major function of the controlled vocabulary is to bring the language of the searcher into coincidence with the language of the indexer.

The controlled vocabulary exists primarily to assist the searcher in the information retrieval process. It prevents the dispersion of related subject matter, reduces possible ambiguities among terms, tells the searcher how various topics have been indexed, and provides enough structure to facilitate the conduct of generic searches, thereby alleviating the problems faced by the searcher in trying to identify all possible terms that might have been used to represent a particular broad subject. A controlled vocabulary may take the form of a classification scheme, a list of subject headings, or a thesaurus of descriptors.

One of the major functions of the controlled vocabulary is to control synonyms. This specifies which of several synonymous expressions is to be used by indexers and searchers and avoids the separation of identical subject matter under different terms in the system. Such control is achieved simply by choosing one of the possible alternatives as the "preferred term," and referring to it -see or use- from the variants to approach the system. If the vocabulary is well constructed, it brings together terms that are hierarchically related in a formal genus-species relationship, and it also reveals semantic relationships across hierarchies. The major functions of the controlled vocabulary are

- To provide for consistent representation of subject matter by avoiding subject dispersion at indexing and searching by control of synonyms, near synonyms & quasi synonyms and by differentiation of homographs.

- To facilitate the conduct of generic searches by bringing together in some way terms that are semantically related. (Lancaster (1979)).

Modern information retrieval systems generally make use of vocabulary as a thesaurus. A thesaurus is essentially a limited vocabulary of terms in alphabetical order that can be used in indexing and searching. It provides control over synonyms, distinguishes homographs, and brings related terms together. The thesaurus is able to prevent the separation of related material under synonymous terms, to distinguish homographs, and to give the searcher positive assistance in the conduct of a comprehensive search in a particular subject area. But the thesaurus does not solve all the language problems that may occur in an information retrieval system. Thus the indexer assigns a group of index terms ("descriptor") to a document but does not indicate the relationships among them. In a searching system, this phenomenon can cause irrelevant items to be retrieved through ambiguous or spurious relationships among terms. For example, in a report discussing the manufacture of electronic components, one of the operations involves the welding of aluminium, another the cleaning of copper by ultrasonic. The indexer assigns the terms in the indexing of a single report as aluminium; copper; welding; cleaning; ultrasonic. This report would be retrieved when there is a request for information on the welding of copper, the cleaning of aluminium, and ultrasonic welding. These are false co-ordinations between terms. To reduce this kind of problem, the possible way is the use of some form of subheading using one term as a subdivision of another, for example aluminium/welding. In general, these problems are theoretical. So, words are precombined as index terms. PLANT may be ambiguous, but when the word is used with the term STEEL, the ambiguity disappears in searching a retrieval system.

When talking about controlled vocabulary, indexers assign only approved terms within the authorized vocabulary to enter the documents in the system. The thesaurus may contain non-approved terms but the user is referred to a more approved term by a "See" reference. To facilitate retrieval of groups of similar documents, Lancaster[1] recommends the grouping together in the thesaurus of conceptually related terms either by a classification scheme or by using "see also" cross references in an alphabetically arranged list.

In free-text indexing, the index file may consist of at least two components, an accession or control file and a term or descriptor file. The accession file is arranged sequentially by an accession or control number of each document. Each item in the accession file contains the full bibliographic information with a descriptive annotation or abstract of the document. The descriptor or term file consists of the indexed subject descriptor items. For example, if we were interested in the solubility of potassium sulphide, we would match the accession number for potassium sulphide with that of solubility. When the

[1] **Lancaster, F.Wilfred. "Information Retrieval Systems", New York: John Wiley & Sons, Inc., 1968. 222p.**

matches occur, those documents contain data on the solubility of potassium sulphide.

When a request for information is received, the searcher must analyze the query as previously indicated and match its key words against the index file. If there is no match between the terms of the query and the index, it must be assumed the requested information is not stored in the system. A document's index terms also permit selective dissemination of information. Therefore, the index file is a key component of an information system. To provide effective searching capability from an alphabetical index file or thesaurus, a type of classification system is built in by means of cross-references. The thesaurus of Engineering and Scientific Terms (1967)[2] which is a controlled list of index terms uses five cross reference terms.

- USE - When a listed term is followed by the instruction to "use" another term, this indicates that the referred to term(s) should be used instead of the listed term.

- UF - The "used for" reference is the reverse of the "use" reference.

- NT - The narrower terms listed under indexing or more specific inquiries.

[2] Thesaurus of Engineering and scientific Terms, prepared for the United State Department of Defence by the office of Naval Research Project Lex in joint operation with Engineering Joint Council, 1967, 690p.

- * BT - The broader term is used to refer from a term representing a member of a class or classes to any term(s) representing that class or classes.

For each broader term reference must be provided with a corresponding narrower term reference.

- * RT - A term preceded by this notation is used to retrieve other terms that are closely related. The related term cross reference can be used to change the scope of specificity.

Some thesaurus use different code symbols such as See = Use; See also = Related Term; includes or specific to = Narrower Term; Generic to = Broader term.

The thesaurus should be a helpful guide to the user. Related terms (RT), Broader Terms (BT), Narrower Terms (NT) are listed under an index term to serve as a reminder of other terms which may be of relevance. The index file is essential to retrieval of both input and output depending on the indexing operation. At the input stage, the indexing analytical operation assigns descriptive labels to documents; at the output the indexing analytical operation assigns descriptive terms to the substantive content of the inquiry.

If the permitted indexing words are too general, then we may index and search for a document by a generic term, when the real target of the query is a more specific topic. We may retrieve the document but find it to be irrelevant waste. On the other hand, if our indexing words are more specific than our queries, there may be many misses, because we do not follow up all the specific words that may be relevant to a generic query. It is important to match the specificity of indexing words to the kind of query that the system has to meet. The thesaurus thus provides more access points for the information search. This feature is particularly apparent when the number of index entry terms is restricted, and the thesaurus lists a whole group of words that each term is 'used for'.

In free-text indexing, whenever new records are added to the database, the computer updates the indexes to that database. Terms can be defined in several ways to meet the requirements of the database. A term might be defined as one word, as a phrase, or as both. This will become clearer by looking at the index terms, which might be generated from the sample record. Most fields have been indexed word by word. This allows searching to take place on individual words in the title and abstract fields as well as by language, year of publication and accession number. Such a capacity provides a very powerful retrieval mechanism. Subject searching is not solely reliant on use of assigned index terms in the descriptor field but can draw upon the words that the author has used in the title and the abstract of the article. Using one of the Boolean operators can then link these terms.

In the free-text indexing system, the users can describe their interest in their own words but the user should know what words would an author have used to convey this idea. To control these obstacles the user has to use term in the alphabetical thesaurus showing broader, narrower and related terms so that the thesaurus would suggest words that are semantically related to each concept. Each interest can now be cast into the form of a preliminary profile by chosen words linked in a logical equation by 'or', and 'or not'. In expanding a concept, one result is to produce a set of related words very similar in spelling. Instead of linking these by 'or' it is possible to capture all of the them by truncation. For example, SYNTHE*, where the asterisk means accept any position, including an end-of-word space. Asterisks may also be used as prefixes, so *SYNTHE* would capture PHOTOSYNTHETIC and BIOSYNTHESIS. Truncation can lead to the recall of irrelevant material. For example, *ACID* will capture not only the related terms as PERACID, ANTACID and ACIDITY, but also the unrelated term as HADACIDIN. To guard against this, the users have to use the indexes. Incidentally the index shows another hazard that may cause references to be missed.

In the controlled vocabulary system, to improve our recall for the request, we can move in one of two directions. Either we can reduce the exhaustivity of the formulation or we can reduce its specificity. We usually reduce specificity by moving up in one of the hierarchies. We can also reduce specificity in more than one category simultaneously, as well. Alternatively, we can broaden our search with the object of improving recall by reducing exhaustivity in the formulation. Exhaustivity of a search strategy is obviously related to the co-ordination level but there may not be a strict one-to-one relationship between exhaustivity and co-ordination level. Obviously, exhaustive formulations are responsible for some recall failures, and non-exhaustive formulations cause precision failures.

Interactivity is one of the great advantages of an on-line system. It depends on indexing terms. An on-line system gives the user the capability of developing a search strategy on a trial and error basis. Basically, the search for information on a particular subject has two major components:

- * The conceptual analysis
- * The translation of the analysis

When information on a particular topic is requested, it can be divided into facets. The conceptual analysis has to be translated into the terms used to represent these concepts in the particular database to be searched. The searcher has entered the vocabulary of the database under these conceptual terms. If these terms are not present, using the cross-reference structure of the thesaurus, the searcher is led to other related terms. The information can be converted into a logical search strategy. It is im-

There are ways in which searching effort or search time can be minimized in many on-line systems using truncation features. To be efficient in machine time as well as user time, it is necessary to identify facets of the search and approach the search by these facets first. It is not always possible to recognize them in a search although it may frequently happen.

Some retrieval systems use weighted term searching in addition to the Boolean approach. The logic of the weighted term search is not different from that of Boolean algebra. In weighted term searching, the first step is the conceptual analysis of the request into its component facets and the expansion of each facet through the selection of appropriate terms from the controlled vocabulary.

In processing of information retrieval, the matching subsystem has direct relationship with the indexing subsystem. It may be controlled vocabulary or free-text indexing. This process can be explained how the index term used with search strategy and how much emphasis for it. The index terms are assigned to the documents to represent their subject matter. The database of index terms may be referred to as the index of the system. This is a device to allow index terms to be matched against those of some search strategies.

Conclusion

Information retrieval systems are almost completely dependent on their indexing subsystems. At the input stage, the indexing analytical operation assigns descriptive terms to documents, at the output the indexing analytical operation assigns descriptive terms to the substantive content of the inquiry. Therefore, the documents should have been assigned to a particular class group defined by a particular index language.

To establish that the index performance has been done perfectly, the recall ratio and precision ratio need to be measured. The precision ratio is a measure of the index's success in screening out non-relevant documents. The recall and precision are adequate and powerful measures and it has been shown that they can be used in the management evaluation of operational systems.

In free-text indexing precision ratio is not high, perhaps, because, in some subject areas terms are not so precise and considerable skill may be needed to device good natural language strategies. Free-text terms are particularly useful when new terminology is being used which may not have entered indexing languages or in subjects where concepts cannot easily be represented by a set of index terms.

The thesaurus must be updated when indexable information cannot be translated into the language of the index. This is done by either expanding the scope of an existing index unit to include the meaning of the indexable information or by establishing a new indexable unit. In both cases the necessary cross-references and other relationships need to be recorded for the new or revised index units.

References

1. HARTELY, R.J... [et.al.].) *ONLINE searching principles and practice.* London: Bowker-Saur, 1990, 73.

2. JAHODA, Gerald. *Information storage and retrieval systems for individual researchers,* New York: John Wiley & Sons, 1970, 27-35.

3. LANCASTER, F.W. *Vocabulary control for information retrieval.* 2nd. Ed. Virginia: Information Resources Press, 1986, 1-8, 13-21, 131-153.

4. LANCASTER, F. W. *Information retrieval systems: characteristics, testing & evaluation.* .2nd Ed. New York: John Wiley & Sons, 1979, 181.

5. LANCASTER, F.W. and Fayen, E.G. *Information retrieval on-line.* LosAngeles, Melville Publishing, 1973, 244-262.

6. SALTON, Gerard and McGill, Michael.J. *Introduction to modern information retrieval.* New York: McGraw Hill: 1983.

7. VICKERY, B.C. *Techniques of information retrieval.* London: Butterworths. 1970.

8. WEISMAN, Herman.M. *Information systems, services, and centres.* NewYork: Becker & Hayes, 1972, 62-81.