DESIDOC Bulletin of Information Technology, Vol. 19, No. 3, May 1999, pp. 23-28

Automatic Document Classification

Babita Jaiswal

Abstract

A computer cannot determine the subject content of a document the way the librarians do. Automatic classification is concerned with the procedures and systems that can make comparison between terms used. Using the computer to provide automatic indexes can only be considered for the automatic production of title indexes or the keyword enhanced index. Classificatory techniques, in fact, can be used for improving the efficiency of these title indexes. This paper attempts to trace solutions to the problems, such as truncation, stemming, etc. Automatic keyword classification is only an attempt. The use of automatically constructed classification is still a subject for investigation.

1. INTRODUCTION

Classification is basically thought of in terms of subject headings, i.e., documents are grouped according to their subject matter as determined by skilled librarians. This is the main function of classification schemes. But it is not a suitable principle on which automatic classification can be based. A computer cannot determine the subject content of a document the way a librarian does. A computer can process words. The principle on which automated classification system is based is statistical, whereas documents can be classified on the basis of the words they contain. Computers are incapable to make association between concepts, which is fundamental to classification theory. However, computers can perform two basic functions with incredible speeds, i.e., count and compare. In order to harness the power of the computer to provide meaningful classification, we must use these two basic functions. These functions can in fact be used to provide what appears to be a degree of intelligence. It is illustrated in the recent developments in artificial intelligence and its application.

To proceed further, it is essential for us to make a distinction between the two terms—*automatic indexing and automatic classification.* Automatic indexing deals with the automation of the process whereby we assign index terms to documents, that is the terms which are used later to retrieve a subject of documents from a file. Automatic classification is concerned with procedures and systems which can make comparison between terms used to index documents.

From this, we draw conclusions concerning the degree of similarity of the key words used to index documents or the degree of similarity of the document themselves. For automated methods in order to replace fully the manual creation and use of classification schemes, implies the following:

- ❑ A classification system for the whole subject range of bibliographic material can be automatically devised;

- ❑ Automated procedure which correctly allocates any document to the classes provided by the scheme created; and
- ❑ System fully satisfies the requirement of the user to find material on a specific subject and related subjects easily and quickly.

There are two main areas of application of classification methods in information retrieval, viz.,

- ❑ Considering the relationship between documents (document clustering); and
- ❑ Considering the relationship between terms (keyword clustering).

2. KEYWORD ENHANCED TITLE INDEXES

The simple technique using the computer to provide automatic indexes can be considered to be those which deal with the automatic production of title indexes or keyword enhanced title indexes. keyword indexing system, though not a new method, is based on the principle that the title of scientific document represents its contents. Each significant word in the title serves as a keyword in turn when the entries are generated by a compute;, the computer moves the title literally in such a way that a significant word for a particular entry always appears either on the extreme left hand side or in the centre of the column. Non-significant words are automatically omitted. The efficiency of the keyword indexing system depends on the title of a document, i.e., if the title reveals the content fully the system also reveals it fully. The title index provides current awareness to some extent, but for subject retrieval devices there are so many limitations. Some of the limitations are:

- ✶ Dependence on title for indicating the contents of the document,
- ✶ Use of natural language rather than controlled vocabulary, and
- ✶ Lack of reference system for ensuring proper unity of the index.

Precision and recall from these techniques alone helps a great deal to overcome these .limitations.

3. AUTOMATIC DOCUMENT CLASSIFICATION (WORD FREQUENCY)

For improving the efficiency of these title indexes, classification techniques can be used. But any classification system, based upon a preconceived and logically organised set of categories, becomes outmoded and inadequate very soon as a result of dynamic growth of knowledge. What is needed is an empirically derived dynamic system of classification wherein a set of documents at a particular point in time and capable of being recognised with a minimum of duplicated effort is made possible by the use of modern computer technology. In order to apply classification automatically, it is important that we first establish how we can construct representation of a document capable of computer manipulation. It is generally assumed that the more frequently a word is used in a document, the more likely it is that the word is a significant indicator of the subject matter. While describing the frequency of a word the first step is that the document is translated into machine-readable code. A computer program develops a machine listing of all words in the document, arranged by frequency of occurrence and alphabetically within frequency. Functional words, such as articles conjunctions, prepositions, etc. are excluded. Depending upon the degree of sophistication the derived words that have the same stem can be counted either as repetitions of the stem or as separate words. Statistical processing of the document produces a list of words arranged by frequency of occurrence, from which index terms can be selected automatically according to pre-specified rules.

3.1 Problems

There are some problems, when we apply this method, viz.,

- ♦ Application of this rule will invariably result in the selection of a large number of index terms, many of which will have very little retrieval value.
- ♦ Computer, computers, computing, these are related words but computer will treat these

words separately. This problem can be seen if we use online databases.

- Use of synonyms is also a big problem. To introduce a style in writing authors do not use same word they use different words for similar concepts. So it is necessary for the computers to identify similar and related terms.

3.2 Solutions

3.2.1 Need to use more refined terms

There is a need to use more refined technique than simple word frequency counts, if we are to use the words contained in a document as a basis for determining its subject. No need to select words according to the frequency of occurrence in the text, but according to the relative frequency of occurrence of normalised terms. Words belonging to the closed set of syntactic or grammatical words are eliminated from consideration. It is also necessary that the selected terms enable relevant documents to be retrieved in response to a request.

3.2.2 Truncation and stemming

It is a solution for the second problem. Truncation allows the searcher to input an arbitrary symbol at the end of part of a sought after term to be used to indicate that any sequence of characters occurring in this position should constitute a match if preceded by the input term. It is possible to search on parts of words by the principle of truncation. Four types of truncation can be recognised.

1. **Right Truncation** is ignoring the ending of a word. Thè truncation CRACK will for example retrieve documents containing CRACK, CRACKS, CRACKING, CRACKED.
2. **Left Truncation** is ignoring the beginning of a word. The truncation mycin, for example, is likely to retrieve documents discussing a large group of antibiotics.
3. **Simultaneous left and right truncation**
4. **Infix Truncation**, i.e., specifying the beginning and end of a word but leaving the middle unspecified.

Right truncation is the most useful for most application and it is a powerful device for searching on large groups of related words.

One major fault of truncation is that it may also bring in same words that have no relevance to a particular request. It must be used with care.

Stemming is a technique which eliminates common inflectional suffixes and prefixes. It is able to reduce words which are common.

3.3.3 Use of Thesaurus

The solution of the third problem is the use of a thesaurus. For controlling synonyms we can use thesaurus. But it is the product of considerable intellectual effort. To automate this activity there is need to produce a system which could use the computer to generate keyword classification.

4. CLUSTERING OF DOCUMENTS

Several attempts have been made at automatic clustering in the area of documents. Some such methods are single linkage, complete linkage, group average and word clustering. One important among all hierarchic methods is single link method. It is based on dissimilarity coefficient. The output from this method is a tree type structure in which clusters are represented by nodes on the tree. A lot of researches have been done in this area, but except of Karan Spark Jones and Salton all the experiments lacked proper evaluation of those automatic classification.

5. SMART

The SMART system works on the principle that the use of many techniques are likely to be more efficient than one. It is a fully automatic system for analysis, search and retrieval of natural language text.

The SMART project was initiated in 1961 at Harvard University. Its aim was to design and implement on a computer, a fully automatic document retrieval system which is capable of processing documents and search requests and or retrieving those documents most similar to the corresponding search requests. From its inception, the system was designed both as a retrieval tool and as a vehicle for evaluating the effectiveness of automatic search and analysis technique. The initial version of the SMART system consisted of three main parts:

(1) An automatic content description system which used a variety of text processing devices supplemented by stored dictionaries to assign a set of automatically generated content identifiers to each stored information item and to each incoming search request.

(2) A supervisory or monitoring system which could process the query and document contents and by suitable matching operations. It could supply to the users *references* to those documents whose content vectors appeared to be most similar to the corresponding query vectors.

(3) An evaluation system which provided formal measurements of system's effectiveness in terms of satisfaction of user needs.

For actual processing in the SMART system, user indicates the particular processing options which are to be operative at a given time, i.e., whether thesaurus, concept hierarchy, syntactic processing, etc:, are to be incorporated into the current run. The SMART system accepts those input instructions, together with document collection, user queries and stored dictionaries. The monitoring system then calls on the appropriate subroutines at the required time.

In the SMART content, the main evaluation viewpoint taker is the user, and the overriding criterion of system effectiveness is the ability of the system to satisfy the user's information need.

5.1 SMART-MEDLARS Comparison

One of the aims of SMART project has been the comparison of fully automatic text processing systems with the manual indexing system that are widely used in operational information retrieval environment. The MEDLARS system has been operating for many years on large databases of several hundred thousand document citations in the medical area, using a controlled indexing vocabulary applied by trained indexers and searchers. The SMART system was tested in a comparison with the MEDLARS in a series of experiments which led Satton to the confident conclusion that the SMART system worked so well that no technical justification appears to exist for maintaining controlled manual indexing in operational *retrieval environment.*

6. KEYWORD CLASSIFICATION

keyword classification is the automatic construction of classification schemes through an examination of keywords used in indexing documents. Its primary objective is not to cluster the document collection but to derive grouping of keywords. keyword classification can be used to characterise the document in a collection as an aid to retrieval. It is required as a device for providing characterisations for all the documents in a collection. The use of a keyword classification itself is familiar to any one who has worked in information retrieval. If the characterisation of a document or a request is by a single keyword or term list, a keyword thesaurus is intended as a device that permits

Table 1. Simplified summary of SMART processing option

Dictionary lookup	User of alphabetic word dictionary		
Concept hierarchy	Concept expansion by *moving upward in* hierarchy	Expansion by *addition of term on* same level using *the theorises*	Expansion by downward movement in hierarchy
Sentence co-occurrences	Extraction of significant sentences		
Syntactic processing	Use of criterion phrase dictionary for matching purposes		
Term associated	Use of term correlations for expansion of document identification, statistical phrase and term cluster formation		
Document association	Use of correlation between document identification and information request		

matching between requests and documents that do not actually contain the same word. Keyword classification is in fact a vocabulary normalisation device rather than means.

8. AUTOMATIC KEYWORD CLASSIFICATION

Automatic keyword classification depends on the hypothesis of co-occurrence where classes consisting of words with a tendency to co-occur in documents will permit satisfactory sub-situations, i.e., ones that promote the retrieval of relevant documents. If two words co-occur in a document, they can be regarded as alternative representative of the topic. If they co-occur in every document in which either occurs, they can be presumed to have a very strong topic connection.

Now, the question comes whether we can obtain a better retrieval performance with a keyword classification than with the keywords alone or whether we can construct such a classification automatically? There are some difficulties in automatic keyword classification:

* It may be the result of an experiment to the character of the classification used, or to the fact that an automatic procedure was used to generate it.
* We have to consider not only the character of the classification itself but how it is used as a component of an indexing language or a retrieval system.
* There is a general problem of establishing a satisfactory methodology in information retrieval research.

9. THEORY OF CLUMPS

Research in automatic classification for information retrieval is associated with the theory of clumps. The criterion for a clump is that the terms in it should be more strongly associated with each other than with the terms outside the clump. The clumps are based entirely on the terms chosen to index the documents. Sometimes it looks like conventional classificatory groupings. For example, grammar, parts of speech, adjective, proposition, phrase, phrase marker, tense, stem, syntax, look very much like the sort of groups one would construct intellectually except that title does not include terms such as 'noun', 'verb', etc., because they do not occur in the indexing of this particular set of documents. The application of clumping techniques to data of different kinds gave varied results. Clump Classification, if derived from similarity matrices, works very well compared with other types of classification. Theory of clump has not yet been wholly satisfactory in itself, because:

- ❑ It explicated our intuitive idea of a class as a set of things that are somewhat loosely organised by family resemblance; and
- ❑ It did not require perfect data, or depend on over simple geometrical models of the kind that treats clauses as footballs.

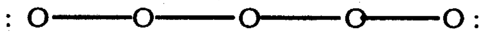
The theory of clumps has just stimulated research in automatic classification in a way.

10. STRINGS, STAR AND CLIQUES

The test results from using the clumping technique proved to be disappointing but work continued to refine the model for clumps further by using statistical and probabilistic techniques. The other form of relationship identical are strings, star and cliques.

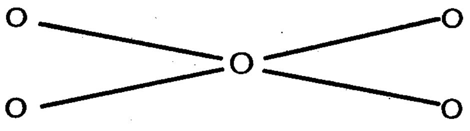
Strings are the simplest form. They are constructed by starting from one specific keyword and finding a second keyword strongly connected with it. The process is repeated with the second keyword. A maximum length of seven terms was permitted in the construction of strings.

Strings



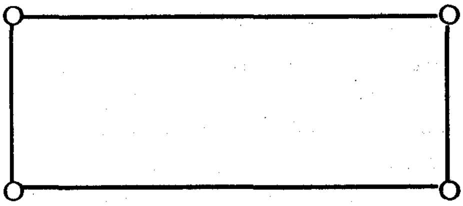
Stars are taken as their starting point with a single central keyword, and other keywords connected with this central keyword are included on the basis of their strong allocation with this term.

Stars



Cliques are obtained by identification of a set of fully interconnected keywords. Techniques used in the evaluation were complex and results are difficult to interpret. However, the conclusion is that information retrieval system can be enhanced through the use of keyword classification clearly.

Cliques



11. CONCLUSION

Though research is taking place in the various fields of automatic classification, it has been seen that automatic classification is not in a position to replace manual classification. The use of automatically constructed classification is still a subject for investigation. The reason behind this is simple. Although so many techniques have been developed but successful implementation of such techniques has not yet been made possible. It is because there is no complete theory of knowledge representation. So automatic classification still remained a goal rather than a reality. But it is also true that automatic classification is an interesting phenomenon in the field of classification. Therefore, it can be said that the future of classification will be just as interesting as its long history.

REFERENCES

1. Bakewell, KGB. The SMART System 1961-1976, Experiments in dynamic document processing. Encyclopaedia of library and information science, vol. 28, pp.1-36.
2. Bosko, Harold. Automatic language processing. John Wiley and Sons, New York, 1968.
3. Dhyani, Pushpa. Library classification: Theory and principles. Mishra, New Delhi, 1998.
4. Foskett, AC. Subject approach to information, Ed. 4. Clive Rustey, London, 1982.
5. Jones, Karen Spark. Theory of clumps. Encyclopaedia of library and information science, Vol. 5. pp.208-24.
6. Narcella, Rita & Nemton, Robert. A new manual of classification. Jaico, Mumbai, 1997.
7. Rijsbergen, Van. Information retrieval. Butterworth, London, 1975.

__Ms Babita Jaiswal__ is presently associated with the Department of Library and Information Science, Lucknow University, Lucknow.