CHAPTER 2

THE INDEXING FUNCTION

1. Introduction

The nature of indexing is examined in this chapter with emphasis on its relation to the document retrieval process. In this context, document indexing may be viewed as a nonreversible (information lossy) transformation from the natural language to an artificial language (the index language), suitable for retrieval purposes. The index transformation is designed in general to accomplish two objectives: it serves on the one hand to trade amount or quantity of information for search speed (indexing produces an information compression), and on the other hand the index transformation serves a language normalization function. Since indexing produces information compression, the index representation of a document can be stored and manipulated with greater facility than a representation of the original text. The index transformation also serves as a language normalization function in the sense that both vocabulary and structure in the index language can be controlled, whereas in the natural language they cannot. The general goal of the indexing function in the context of document retrieval, then, is to provide a compact representation of the information content of source documents (or arbitrary segments of natural language texts) in a controlled format.

2. Manual Indexing

The increasing emphasis which has been placed in recent years on coping with the rising flow of technological literature has stimulated study of the traditional goals and procedures of the various activities of documentation and library facilities.[1] The range of services offered by such facilities is too broad to be considered here. Since subject classification and document indexing play a primary role in facilitating the accessibility of recorded knowledge, these facets of documentation have been examined with particular care.[2,3,4,5] Such activities are directly concerned with the document retrieval problem, and thus are germane to this discussion; on the other hand, the present objective is to consider mechanized systems for literature searching. The theory and methodology of manual indexing must, therefore, largely be ignored. In so far as manual indexing is considered as an input process for mechanized search and retrieval systems, its goals are similar to those of automatic indexing. In this sense, the limitations of current techniques for linguistic data processing normally dictate the form and structure of the index language. Thus the scope of both manual and automatic index transformations for mechanized retrieval are largely the same and the distinction between the two is not critical for an examination of the structure of index representations of information content suitable for mechanized processing.

3. Automatic Indexing

The goal of automatic indexing is to develop a set of computer-

based linguistic analysis procedures which provide an effective representation of the information content of source documents without manual intervention at any stage of the process. Information is conveyed in the natural language by the variety of semantic and structural constraints implicit in the language. Machine indexing techniques all depend, in effect, on the automatic recognition of some set of the information carrying elements of the natural language, and on the representation of these elements in a formal structure. In general, the processes of automatic content analysis can be classified according to whether they are statistically, semantically, or syntactically based. A discussion of each classification follows.

A. The Statistical Approach

A natural starting point for statistical content analysis consists in assuming that meaning is principally carried by the words used in a document.[6] Under this assumption a suitable index transformation consists in mapping a document into an unordered set of significant content bearing words extracted from it. A variety of statistical techniques have been proposed and investigated for determining the most suitable set of content words (keywords) to be used for the encoding.[7,8] Typically, such techniques generate a frequency count of word types (ignoring most function words) and then invoke some frequency sensitive selection process to produce the document index image. Such procedures can, of course, be extended in theory to the detection and counting of word pairs, triples, etc.[9]

There exist obvious problems to such extensions on the practical level.

An important defect in such an index transformation lies in the fact that the structure of the index image provides no facility for representing the semantic associations which exist between distinct word types in the natural language. One proposal for dealing with such associations on a statistical basis consists in assuming that they can be derived a posteriori from a set of index images characterizing a document collection in some given subject area. Thus one can assume, for example, that terms which co-occur in the sentences of a given document, or in the documents of a given collection, more frequently than the average are, in fact, semantically as well as statistically related.[10,11,12] In the formal associative model, it is possible to account for key word associations of higher order than the first and in addition to use these associations to influence query-document matching procedures. In such a system, a document is represented by its keyword set and additionally by the statistical properties of the representations of all other documents in the collection.

B. Semantic Techniques

An important alternative to the statistical associative process consists in providing a specific semantic model in the index transformation directly. The indexing function may then be implemented by a thesaurus mapping containing a pre-defined set of semantic associations. A thesaurus transformation may be defined as a many to many mapping from recognizable word types or phrases to thesaurus

categories or concept codes. Thus a set of semantically associated natural language terms comprised of synonyms, for example, can be mapped into a single element in the index language; or a single natural language term which has several connotations can be identified with a set of elements in the index language (homonyms might be treated in this manner). Figure 2.1 provides an illustration by means of an excerpt from the SMART system thesaurus.

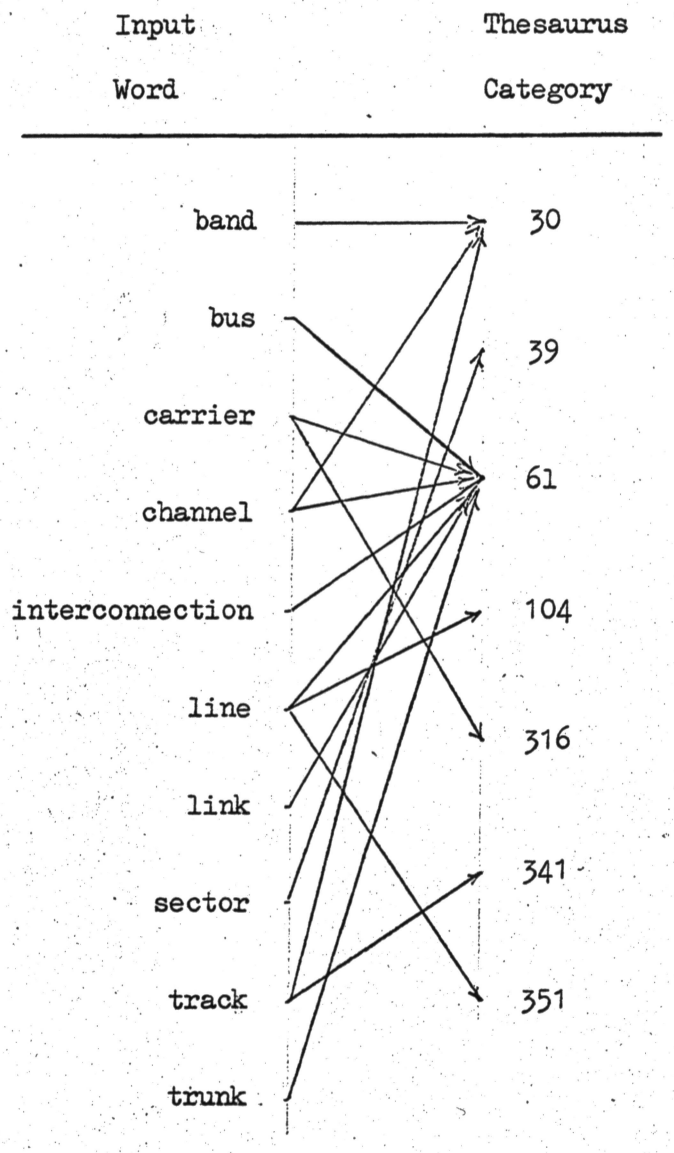
The notion of a semantically based transformation on a set of recognizable (by machine) linguistic features (word or stem types, phrases, etc.) can be generalized to include a variety of the associations which such elements possess.[13] The index transformation may be described in this case by considering a multi-stage mapping. The first step consists in mapping the document into the set of basic elements which describe it, e.g. into the set of word types it contains. The second step is a transformation from these elements into a space of synonymous term groups i.e. into thesaurus categories. (The thesaurus mapping described above consists in applying these two basic transformations.) Additional transformation stages may also be defined. Thus generic (inclusion) relations exist among semantic elements and these may be used to define a set of hierarchies. A number of transformation can be defined based on a set of such relations; thus a term which includes or which is included by a given term may be added to or may replace the related term in the document image. The index image of a document, therefore, can be modified to contain terms which are generically related to those detected, but not explicitly present in the input text. Relations among index terms other than

inclusion relations are also of interest, e.g. cause-effect, process-products, etc. and these can be used to define additional transformations. For example, the class of elements denoting processes can be identified and the corresponding products listed. A document image containing a process term may then be modified to include the associated product term and vice versa.

In summary it is possible, then, to consider semantic index transformations which include a variety of term associations such that in principle a multiplicity of index representations can be produced based on the same set of machine recognizable linguistic features. The problem with such transformations, in general, is that a large number of a priori semantic associations are possible among the index terms describing a given document. The correct associations are dependent on the context in which the terms are used so that a context free encoding such as is generally produced by machine processing does not necessarily improve the accuracy of the index representation of information content.

C. Syntactic Techniques

In general both the statistical and semantic procedures discussed above ignore the information carried by the structural constraints of the natural language. It is possible, however, to incorporate a number of syntactic recognition features into automatic indexing algorithms. One obvious use of this kind of information is stem detection, i.e. recognition of the intrinsic association of the various morphological forms of a given word. Stem detection is readily



Excerpt From the SMART Thesaurus

Figure 2.1

applicable to both statistical and semantic processing since the meaning of a word is generally invariant over its morphological variants. Much more ambitious syntactic processing procedures are also under investigation, including the use of fully automatic syntactic analysis. A full sentence by sentence syntactic analysis could, for example, provide explicit dependency relations among the various semantic elements of a sentence, and could be used for phrase recognition or for the recognition of structurally constrained associations among semantic terms. At the present time it is not clear whether the complexity required for the recognition of complex structural constraints is justified in terms of the additional information extracted thereby.

4. The Structure of Index Representations

The index transformation represents a mapping from the natural language of the source text to the target or index language. The index image of a source document is thus a representation of the content of the document in this target language. The most commonly used index language structure is the description list, or property vector, in which the index image consists of a list of those properties of a finite set which characterize the document. Index images of this type are used, for example, in Uniterm systems where the document representation is an unordered set of keywords (descriptors, uniterms, etc.). If the property set is ordered, for example, by a 1 to 1 mapping to the set of integers, the index image may be encoded as a binary vector. A more general representation of the same type allows for a quantization of the

value, or degree, to which each attribute pertains to the document by associating a scalar with each attribute. In this case the index images can be encoded as numeric rather than binary description vectors in the attribute space. Table 2.1 illustrates a typical keyword description derived by statistical analysis (from Booth[16]) in which the relative frequency of the 15 most frequent non-common word stem types from a sample document are shown. This analysis can be used to establish a property set index image (by employing a frequency sensitive selection procedure), a binary description vector, or a numeric description vector incorporating relative frequency information. Symbolic examples of each of these are illustrated in Figure 2.2.

A property list description does not allow for a direct representation of any relations among the various attributes, unless these are specifically identified in the attribute space. Since information in the natural language is conveyed by semantic referents (words, phrases) and by the relations indicated among the referents (syntax and context), index languages capable of explicitly representing relations among attributes have been investigated. A variety of such structures have been studied,[17] including tree and graph representations. A syntactic dependency tree, for example, can represent a natural language sentence by associating its nodes with the semantic values of the words they represent, and its branches with direct syntactic dependency. An example (from Sussenguth[18]) is illustrated in Figure 2.3. While such index structures are capable of more precise modeling of the information carrying elements of the

Word Stem Type	Relative Frequency
term	.53
document	.28
request	.20
index	.11
profile	.11
association	.098
generation	.086
number	.069
related	.045
information	.045
collection	.045
system	.041
relevance	.037
expanded	.037
thin	.033

Word Stem Frequency List of a Sample Document[x]

Table 2.1

[x] The document characterized is reference 12.

Keyword Set: $\{A,B,C,D,E,F,G,H\}$

a) The Property Space

Keyword	Relative Frequency (Significance)
A	1
B	3
E	10
H	5

b) Tabular Document Representation

$$d = \{E,H,B,A\}$$

c) Document Image as a Set

$$\bar{d} = (1,1,0,0,1,0,0,1)$$

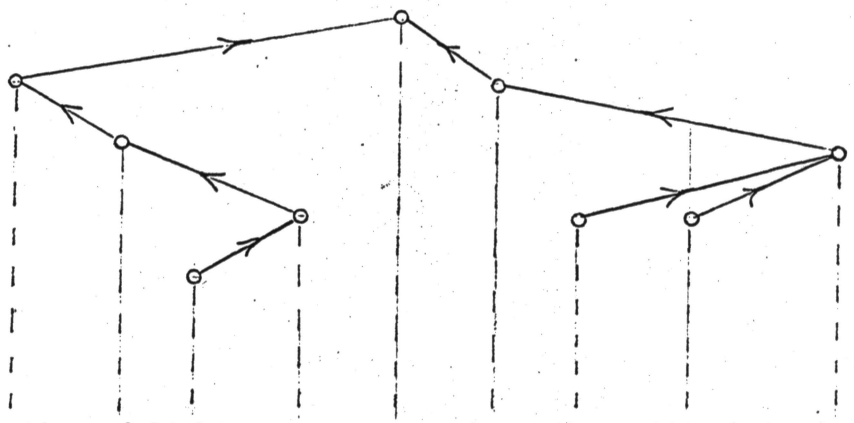
d) Document Image as a Binary Property Vector (Keywords in lexical order)

$$\bar{d} = (1,3,0,0,10,0,0,5)$$

e) Document Image as a Numeric Property Vector (Keywords in lexical order)

Alternative Property Space Index Representations

Figure 2.2



Radiation of light waves occurs above the critical temperature.

Syntactic Dependency Tree

Figure 2.3

natural language, they necessarily involve text processing algorithms more costly in terms of complexity and time, in addition to requiring more storage. The degree to which such index languages justify their additional cost in terms of increased retrieval performance is at the present time an open question.

5. Optimizing the Index Transformation

The two major aspects of automatic indexing algorithms are the mechanical detection of information carrying elements of the natural language and the representation of these elements in the index language. To be useful, the recognition procedures must be applicable to a sufficiently wide class of documents such as will be found in the literature of some subject specialty. Due to the variations in usage

and linguistic style, and due to the difficulties of extracting contextual information, any set of properties chosen to encode the information content of documents or search requests in a given field must reflect statistical approximations over the usage of the detected features. Such a statistical basis is clearly evident in the statistical association indexing model discussed earlier, where it forms an explicit part of the index representation. In various other indexing schemes such as manual descriptor indexing, or in mechanized thesaurus indexing, the statistical approximations are, in effect, hidden in the decision rules incorporated in the index transformation. This necessary statistical basis for document content encoding is emphasized because of its significance in terms of the problems of generating, maintaining, and evaluating indexing schemes.

Consider as a concrete example the indexing model specifically assumed in this report. The semantic associations incorporated in the thesaurus mapping from word stems into thesaurus or concept categories can be established on an ad hoc basis, reflecting individual or collective value judgments. It is possible, however, to subject these value judgments to experimental verification. Assume, for example, that a given set of natural language terms (words, phrases, etc.) is mapped into a single attribute of the index space, i.e. all the elements of the set have been judged to be sufficiently associated so as to be treated as a unit in the index language. It is possible, then, to examine the occurrence of this

property in some sample set of document index representations, and to manually determine from the context of each occurrence in the source text whether the semantic value assumed in the index transformation, in fact, agrees with that found in the document. The degree to which actual usage conforms to the associations assumed in the model can provide confirmation or suggest changes.

One possibility is to incorporate such statistical evidence directly into the thesaurus transformation by use of a weighting scheme. Consider the mapping shown in figure 2.1. The term "channel" maps into two categories, one of which, category 30, is associated with magnetic disk storage, while the other, category 61, is associated with information transmission. On the basis of the statistics of a collection of documents the a priori probabilities of each of these usages can be estimated. Assume that the category 30 context occurs with relative frequency α and that the category 61 usage occurs with relative frequency $1-\alpha$. The contribution of n occurrences of "channel" in a document will then contribute an amount $k \cdot n \cdot \alpha$ to the resultant weight of category 30 and $k \cdot n \cdot (\alpha - 1)$ to category 61 (where k is an arbitrary scaling constant). In any event, assuming that such a procedure or its equivalent is carried over a sufficiently large sample of source text to produce statistically significant correlation of the various associations incorporated into the index transformation, the index image of a particular document is still at best a good approximation of what could be produced manually by applying a similar set of context dependent rules. In other words the noise introduced by

the statistical approximations necessary in a context independent framework must necessarily distort the characterization of the document's content.

This suggests that there are essentially two alternatives to improving an already statistically optimized index transformation. One method clearly involves the incorporation of context dependent recognition procedures into the content detection process. In some sense, this is approximated by encoding larger segments of the natural language text, e.g. phrases instead of words, or sentences instead of phrases. Alternatively, context dependence can be introduced by multi-level recognition procedures in which the decision rules are altered by global interpretation of a context free encoding, thereby producing a second context dependent index representation.

Consider again a thesaurus transformation of the type illustrated in Figure 2.1. Assuming that all ambiguous input terms (terms which map into more than one thesaurus category) are mapped with statistically derived weights as described above, one can expect that the correct context will be reinforced over all the term encodings characterizing the document, whereas the incorrect ones will not. The term "channel", mapped as shown in Figure 2.1, is initially associated with two alternative contexts. After the entire initial encoding is completed, it should be possible to derive a total score for context "magnetic disk" versus the context "information transmission" by comparing the total weights of all

categories associated with each of these concepts. If such an evaluation indicates that one of these alternatives is much more likely than the other, the individual term mappings can be reweighted conditioned on this finding. Thus the weight of category 30 due to occurrences of "channel" might be increased from $k \cdot n \cdot \alpha$ to $k \cdot n$ if the total document encoding indicated that "magnetic storage" was more probable than "information transmission." This second encoding is context dependent in the sense that a global interpretation of the overall original index image has been used to modify the term transformation rules.

The second method of possible optimization which avoids the necessity of adding context dependent complexity to the index transformation, focuses on the index images of individual documents. Consider a document retrieval system in which it is possible to record effectively the reactions of the users to the responses of the system. In the course of a retrieval operation the indexing rules of the system are evaluated by users in the sense that some retrieved documents are accepted and others are rejected. To the extent that these user value judgments reflect the indexing accuracy of the system, the information they represent can be used to modify the index images of the documents in question.

One possible means of implementing such a scheme is by an index term weight adjustment algorithm. Consider a document $\bar{d}_i$ retrieved in response to a search request with index image $\bar{q}_j$. Define the set of terms common to $\bar{d}_i$ and $\bar{q}_j$ to be T_c. If the user judges $\bar{d}_i$

to be relevant, the weights of all terms contained in T_c are promoted in the index image $\bar{d}_i$; while if a negative relevance judgment occurs, the corresponding term weights are decreased. Over a number of such operations those terms most useful in characterizing the content of document $\bar{d}_i$ will receive increased significance while terms representing out of context encodings should receive no net significance gain.

The efficiency of such a scheme is open to question. Clearly, one would want the changes in the index image of a given document to reflect some significant sample of user value judgments, since each user brings his own individual biases and viewpoints into the picture. The rate of adjustment to the index images of documents for a given topic will then be influenced by the density of search requests for that topic. This implies that the documents of most interest have the highest probability of obtaining an improved index representation, a feature which could be desirable. On the other hand, however, a serious drawback of this process lies in the difficulty of improving a poorly indexed document. Such a document will hardly ever be detected as being relevant to any user's search request so that any evaluations obtained will be negative ones. While special techniques (such as calling for manual intervention) might be employed in this case, it seems clear that the value of such an optimization process will depend critically on how good the initial index transformation is to begin with.

The question of whether effective operational procedures can

be devised to implement the optimization of the index images of particular documents, in accordance with information obtained by subjective evaluation of a user population will require a large sample of experimental evidence such as can only be obtained in an operational system environment. The following chapter, however, offers experimental evidence that similar techniques are useful for search request optimization.

REFERENCES

1. Proceedings of the International Conference on Scientific Information, 2 Vols., Washington, D.C., 1958

2. Perry, J.W., and Kent, A., Tools for Machine Literature Searching, Interscience Publishers, New York, 1958

3. American University, "Machine Indexing: Progress and Problems," Center for Technology and Administration, American University, February 1961

4. Cleverdon, C., "Interim Report on the Test Programme on an Investigation into the Comparitive Efficiency of Indexing Systems," College of Aeronautics, Cranfield

5. Cleverdon, C. and Mills, J., "The Testing of Index Language Devices," Aslib Proceedings, Vol. 15, No.4, 1963

6. Luhn, H.P., "A Statistical Approach to Mechanized Encoding and Searching of Library Information," IBM Journal of Research and Development, Vol. 1, January, and Vol. 4, October 1957

7. Maron, M.E., and Kuhns, J.L., "On Relevance, Probalistic Indexing and Information Retrieval," Journal of the ACM, Vol. 7, No. 3; July 1960

8. Maron, M.E., "Automatic Indexing: An Experimental Inquiry," Journal of the ACM, Vol. 8, No. 3; July 1961

9. Doyle, L.B., "The Micro-statistics of Text," Information Storage and Retrieval, Vol. 1, No. 4, Nov. 1963

10. Giuliano, V.E., and Jones, P.E., "Linear Associative Information Retrieval," Repts. CACL- 1,2,3, Arthur D. Little, Inc. 1962-1963

11. Giuliano, V.E., and Jones, P.E., "Linear Associative Information Retrieval," in Vistas in Information Handling, P. Howerton ed., Washington, D.C.; Spartan Books, 1963

12. Stiles, H.E., "The Association Factor in Information Retrieval, Journal of the ACM, Vol. 8, No. 2; January 1961

13. Salton, G., "The Identification of Document Content- A Problem in Automatic Information Retrieval," Harvard Symposium on Digital Computers and Their Applications, April 1961

14. Kuno, S. and Oettinger, A.G., "Multiple-path Syntactic Analyzer," Proceedings of the IFIP Congress 62, Munich, 1962

15. Salton, G., and Sussenguth, E.H.Jr., "Some Flexible Information Retrieval Systems Using Structure- Matching Procedures," Proceedings of the AFIPS Spring Joint Computer Conference, April 1964

16. Booth, A.D., "Characterizing Documents- A Trial of an Automatic Method," Computers and Automation, Vol. 14, No. 11, Nov. 1965

17. Salton, G. and Sussenguth, E.H., "Some Flexible Information Retrieval Systems Using Structure-Matching Procedures," Proceedings of the AFIPS Spring Joint Computer Conference, April 1964

18. Sussenguth, E.H., "The Sentence Matching Program-Graph," Report ISR-7, National Science Foundation, Harvard Computation Lab, Camb., Mass., June 1964