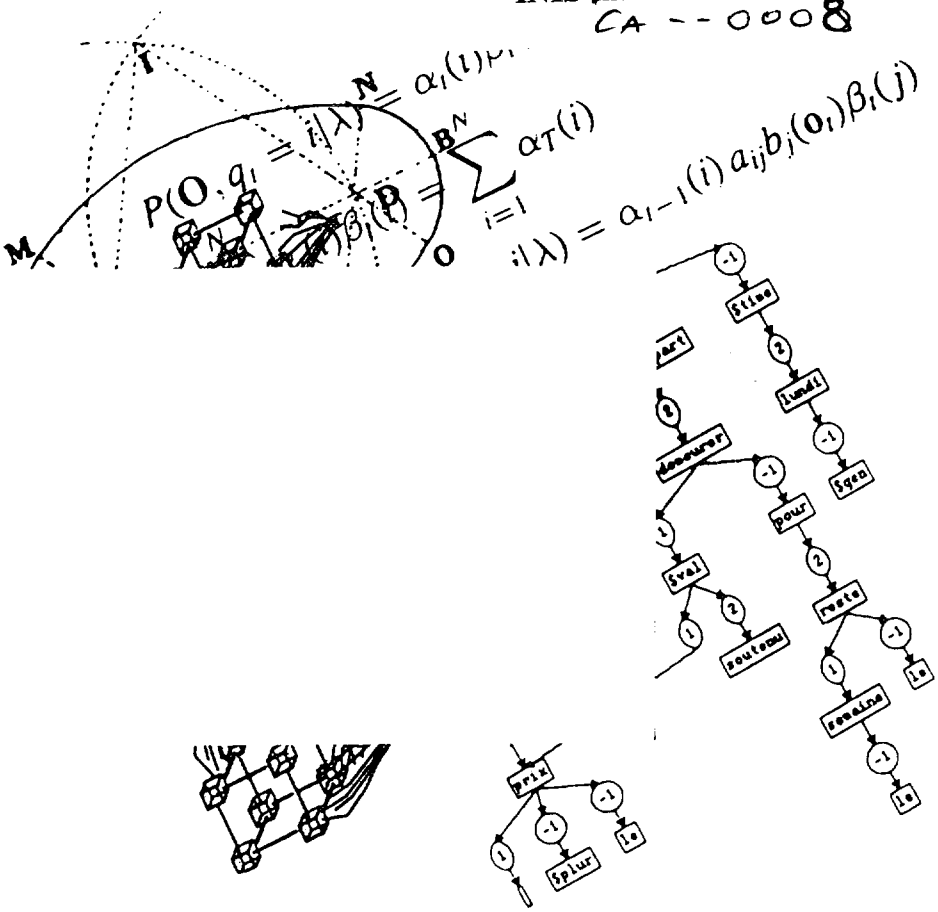

CITI

Centre d'innovation
en technologies de l'information

Centre for Information
Technologies Innovation

Canada

Industry Canada
Centre for Information Technology Innovation (CITI)

INTELLIGENT INDEXING [1]

by

Jennifer Farkas

Laval

June 1991

---

1 Document presented at the Canadian Conference on Electrical and Computer Engineering, Québec, Québec, September 25-27, 1991.

This report was prepared in connection with work carried out by Centre for Information Technology Innovation (CITI) of Industry Cnada. The views expressed in the report are those of the author.

N° catalogue Co28-1/95-1992E
N° ISBN 0-662-20056-X

Abstract

In this paper we discuss the relevance of artificial intelligence to the automatic indexing of natural language text. We decribe the use of domain-specific semantically-based thesauruses and address the problem of creating adequate knowledge bases for intelligent indexing systems. We also discuss the relevance of the Hilbert space l^2 to the compact representation of documents and to the definition of the similarity of natural language texts.

1 The Indexing Problem

Purely quantitative indexing techniques are inadequate for the creation of reliable indexes for non-homogeneous sets of documents. On the other hand, qualitative human indexing is labour-intensive and requires levels of indexing skills that are difficult to achieve and maintain in most management environments. At the Canadian Workplace Automation Research Centre (CWARC) we have begun to apply artificial intelligence techniques to solve this problem and are working on an expert system for document indexing called IndeXpert.

2 An Intelligent Indexing System

IndeXpert is a system for automating the indexing of documents using artificial intelligence techniques. It is a prototype of an interactive bilingual computer-assisted system which uses domain-specific thesauruses to obtain keyword representations of documents. One of the distinguishing features of IndeXpert is that it models expert human knowledge and therefore contributes to a consistent representation of documents. The knowledge base of IndeXpert serves to convert preliminary lists of descriptors into compact representations of documents relative to the classification of terms in domain-specific thesauruses.

Since all document indexing systems use natural language text as their primary input, any content analysis, as Salton and Lesk point out in [12], "will have to include methods for consistent language normalization. One of the most effective ways for providing such a normalization is by means of suitably constructed dictionaries." IndeXpert uses specific types of dictionaries, viz., thesauruses, to give semantic interpretations to term phrases encountered in documents. The motivation behind this technology is that "term relations are identified very often by means of a thesaurus. A thesaurus is a structure where for each term a set of synonymous terms, a set of narrower terms, and a set of related terms are given. Even in a monolingual environment, a thesaurus increases the performance of an information retrieval system." [14]. This feature provides a powerful semantic enhancement to any automatic indexing system, but requires the localization of the system to specific domains for which thesauruses either exist or can be defined. In

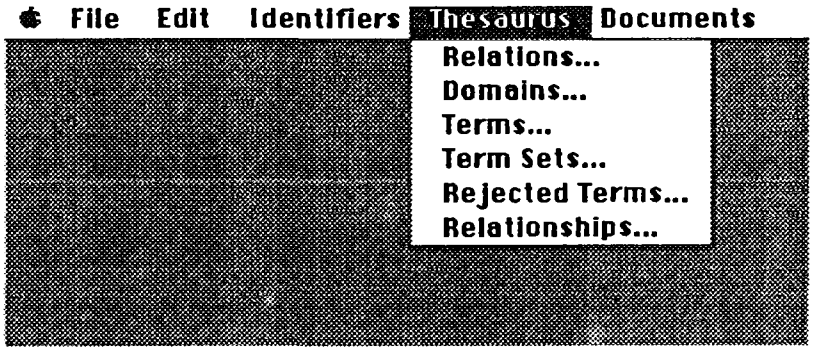

Figure 1: The Thesaurus Component of IndeXpert

the case of IndeXpert, the value of the use of thesauruses is enhanced further through its adherence to the internationally accepted ISO norms which provide standard techniques for the testing of the reliability and completeness of a thesaurus.

3 The Vector Representation

For the purpose of this discussion, we assume familiarity with the subspace of l^2 (cf. [16]) consisting of sequences of real numbers with the property that all but finitely many coordinates are 0. For the purpose of the vector representation of documents we assume as given a fixed thesaurus T of IndeXpert whose terms are well-ordered, i.e.,

$$T = \{t_1, \ldots, t_n\}.$$

We also assume that the documents have been indexed and that IndeXpert has produced corresponding lists of keywords for the documents. We map each document into the subspace of l^2 as follows: Let

$$\begin{aligned} D &= \{d_{j_1}, \ldots, d_{j_p}\} \\ D' &= \{d_{k_1}, \ldots, d_{k_q}\} \end{aligned}$$

be two given indexed documents, where $d_{j_r} = t_{j_r}$ in the well-ordering of

T, etc. Let $< d(t_i) >$ be the sequence of 0's and 1's obtained from D by defining

$$d(t_i) = \begin{cases} 1 & \text{if } d(t_i) = d_i \text{ for some } i \\ 0 & \text{otherwise} \end{cases}$$

Using the inner-product structure on l^2, we let

$$(D, D') = \sum d_i d'_j$$

be the inner product of D and D' and define

$$\| D \| = (\sum |d_i|^2)^{\frac{1}{2}}$$

as the norm of D.

From these measures we can define the cosine of the *angle* θ between the two documents by letting

$$cos(\theta(D, D')) = \frac{(D, D')}{\| D \| \| D' \|}.$$

We say that the documents D and D' are *similar* if the angle θ is small. The definition can of course be made precise and expressed as a function of θ. It should be noted that by using the infinite dimensional space l^2, we are able to give a uniform definition of similiarity that is independent of the size of the particular thesauruses in use at any given time.

In the present context, a *document* denotes the indexing image $IndeXpert(A)$ of a piece of natural language text A under IndeXpert. We say that two pieces of natural language text A and B are *congruent* if

$$cos(\theta(IndeXpert(A), IndeXpert(B))) = 1$$

and are *orthogonal* if

$$cos(\theta(IndeXpert(A), IndeXpert(B))) = 0.$$

3.1 Example 1

Let $D = IndeXpert(A)$ be the indexing image of a document relative to the thesauruses shown in Figure 2, suppose that $D = \{$accounting, agreement, Aladin, Amethyst, annual report, Argument and Decidex, artificial intelligence, artificial intelligence application, authorization$\}$, and let $D' = IndeXpert(B) = D \cup \{$budget$\}$. Then

$$\| D \| = 3 \text{ and } \| D' \| = \sqrt{10}.$$

and

$$cos(\theta(D, D')) = \frac{3}{\sqrt{10}} \approx .95.$$

Hence $\theta(D, D') \approx 18°$. With respect to the semantic similarity measure defined, A and B are considerably closer to being congruent than to being orthogonal.

3.2 Frequency Counts

The given definition of $d(t_i)$ is natural and appropriate for the determination of the congruence and orthogonality of a document. However, the vector space operations on l^2 are virtually irrelevant to these calculations. In particular, the frequency of occurrence of a thesaurus term in a document, which is often taken to be a significant indicator of the relevance of the term in the classification of a document, is not taken into account. We therefore redefine $d(t_i)$ as follows:

$$d(t_i) = \begin{cases} n & \text{if } d(t_i) = d_i \text{ for some } i \\ 0 & \text{otherwise,} \end{cases}$$

where n denotes the number of occurrences of the term t_i in the document being indexed.

Vector addition is now a meaningful operation. If $< d(t_i) >$ and $< d'(t_i) >$ represent two documents A and B, then

$$< d(t_i) + d'(t_i) >$$

represents the document C obtained from A and B by appending B to A.

3.3 Example 2

Let D and D' be the documents discussed in Example 1 and suppose that the term *accounting* occurs twice in document D and the term *budget* occurs four times in document D', with all other thesaurus terms occurring exactly once in each document. Then

$$\| D \| = \sqrt{12} \text{ and } \| D' \| = 5 \text{ and } (D, D') = 10.$$

The cosine of the angle θ between the two document is now given by

$$cos(\theta(D, D')) = \frac{10}{(\sqrt{12})(5)} = \frac{1}{\sqrt{3}} \approx .58.$$

Hence

$$\theta(D, D') \approx 55°.$$

As was to be expected, the underlying documents A and B are less similar with respect to this measure than they were with respect to the previous measure.

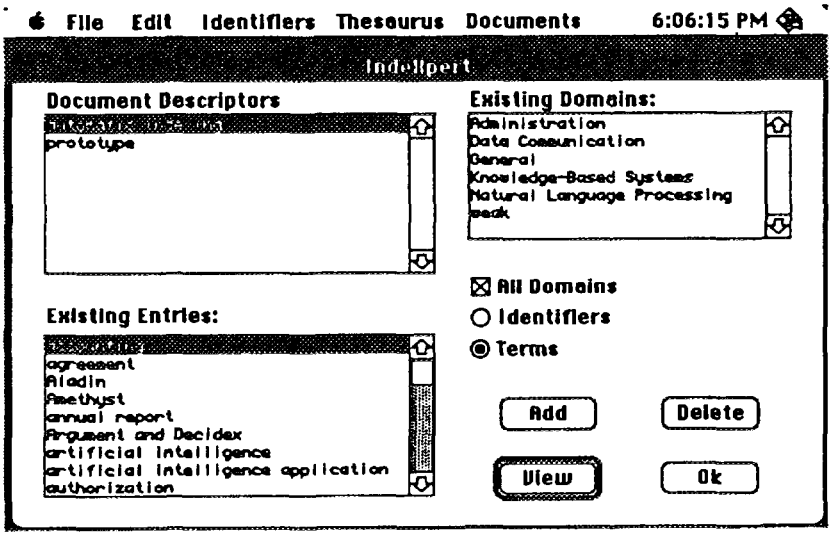

Figure 2: The Indexing Component of IndeXpert

4 Knowledge Bases

IndeXpert provides only a partial solution to the semantic automatic indexing problem since it is well known that purely syntactic techniques cannot capture the contextual meaning of all words in all contexts [11]. The system is therefore designed to formalize human indexing practices for only limited domains. The intelligent component of IndeXpert is its formalization of significant managable fragments of cognitive thinking. The user is provided with the option of a *what-if* capability that allows for the indexing of a given document relative to either all existing domains, or relative to a chosen subset of the domains. For particular documents, more reliable indexing results can be achieved by specifying the domain of discourse. The windows in Figure 2 indicate this capability of IndeXpert. The *Existing Domains* window displays the existing domains, the *Existing Entries* window shows the user the set T of resulting thesaurus terms with respect to which the document can be indexed, and the *Document Descriptor* window displays the keywords for a given document obtained by IndeXpert relative to the chosen domains. A catalogue of proper names in an *Identifier* window provides additional options for the user not covered in the thesauruses.

IndeXpert is intended to provide an improvement over existing systems in several respects:

1. The system uses artificial intelligence technology, which is becoming recognized as an indispensable tool for effective and reliable document indexing. As is pointed out in [11], "syntax by itself cannot resolve the many ambiguities that complicate the content analysis task. Various attempts have been made in the recent past to use syntactic analysis methods for the generation of complex constructions, such as noun and prepositional phrases, that are essential for content identification in various automatic text analysis systems." One of the purposes of IndeXpert is to address this problem.

2. The localization of the indexing problem to specific domains and the use of domain-specific thesauruses increases the reliability, speed, accuracy, and completeness of document retrieval.

3. The use of domain-specific knowledge bases to capture specialized human indexing skills. As pointed out in [17] "the decisions made by

indexers in their selection of descriptors to index the literature represent a large intellectual investment in any one database. Since most indexers have a professional background in the particular discipline whose literature they index, their indexing efforts constitute a collection of expert decisions about the subject content of the literature."

4. Most existing indexing systems are mainframe systems. The fact that IndeXpert is designed to function as a stand-alone system for the PC environment increases the potential user base.

5 Future Work

The current prototype of IndeXpert was developed for the Macintosh environment in the artificial intelligence language Prolog. It reads English- and French language documents, determines the language of a document, chooses the appropriate thesauruses, and automatically indexes the documents. At this point, its knowledge base consists of a variety of generic indexing rules that are thesaurus dependent. The most significant enhancement planned for IndeXpert is the extension of the current rule set to domain-specific rules which capture the working practices of indexing experts. In this way a respectable improvement in the quality of document management in specific domains can be expected to be achieved.

References

[1] G. Biswas, J. C. Bezdek, M. Marques and V. Subramanian, "Knowledge-Assisted Document Retrieval: I. The Natural-Language Interface", *Journal of the American Society for Information Science,* vol. 38 (1987) 83-96.

[2] G. Biswas, J. C. Bezdek, V. Subramanian and M. Marques, "Knowledge-Assisted Document Retrieval: II. The Retrieval Process", *Journal of the American Society for Information Science,* vol. 38 (1987) 97-110.

[3] W. B. Croft and R. H. Thompson, "I^3R: A New Approach to the Design of Document Retrieval Systems", *Journal of the American Society for Information Science,* vol. 38 (1987) 389-404.

[4] J. R. Driscoll et al., "The Operation and Performance of an Artificially Intelligent Keywording System", *Information Processing and Management,* vol. 27 (1991) 43-54.

[5] D. Harman, "An Experimental Study of Factors Important in Document Ranking", in: (Fausto Rabitti, editor), 1986—ACM Conference on Research and Development in Information Retrieval, 8-10 September 1986, Pisa, Italy, 186-191.

[6] H. J. Jeffrey, "Expert Document Retrieval via Semantic Measurement", *Expert Systems with Applications,* vol. 2 (1991) 345-352.

[7] F. W. Lancaster, "Vocabulary Control for Information Retrieval", Information Resources Press, Arlington, Virginia, 1986.

[8] L. C. Malone, J. R. Driscoll and J. W. Pepe, "Modeling the Performance of an Automated Keywording System", *Information Processing and Management,* vol. 27 (1991) 145-151.

[9] "Norme internationale ISO (5964)", Documentation-Principes directeurs pour l'établissement et le développement de thésaurus multilingues, 1985.

[10] G. Salton, "Automatic Text Processing", Addison-Wesley, New York, 1989.

[11] G. Salton, C. Buckley and M. Smith, "On the Application of Syntactic Methodologies in Automatic Text Analysis", *Information Processing and Management,* vol. 26 (1990) 73-92.

[12] G. Salton and M. E. Lesk, "Information Analysis and Dictionary Construction", in: (G. Salton, editor), The SMART Retrieval System, Prentice-Hall Inc. (1971) 115-142.

[13] G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, New York, 1983.

[14] P. Schäuble, "Improving the Effectiveness of Retrieval Systems by Information Structures", *Information Processing and Management,* vol. 25 (1989) 363-376.

[15] L. C. Smith, "Artificial Intelligence and Information Retrieval", in: (Martha E. Williams, editor), *Annual Review of Information Science and Technology,* vol. 22 (1987) 41-77.

[16] A. E. Taylor, "Introduction to Functional Analysis", John Wiley and Sons, New York, 1964.

[17] C. Todeschini and M. P. Farrell, "An Expert System for Quality Control in Bibliographic Databases", *Journal of the American Society for Information Science,* vol. 40 (1989) 1-11.