On the Use of Term Associations in Automatic Information Retrieval

Gerard Salton*

Abstract

It has been recognized that single words extracted from natural language texts are not always useful for the representation of information content. Associated or related terms, and complex content identifiers derived from thesauruses and knowledge bases, or constructed by automatic word grouping techniques, have therefore been proposed for text identification purposes.

The area of associative content analysis and information retrieval is reviewed in this study. The available experimental evidence shows that none of the existing or proposed methodologies are guaranteed to improve retrieval performance in a replicable manner for document collections in different subject areas. The associative techniques are most valuable for restricted environments covering narrow subject areas, or in iterative search situations where user inputs are available to refine previously available query formulations and search output.

1. Introduction

Computers were first used for the processing of natural language texts over 30 years ago. From the beginning it has been recognized that the individual words contained in the texts of written documents could be used in part to provide a representation of document content. At the same time it was generally accepted that certain words, or word sets, would not produce meaningful content identifiers. In particular, some quite broad words, such as the term "computer" used to identify computer science literature, would be useless for distinguishing one document from another. Other very specific terms would be so rare that no single item in a collection might reasonably be described by such a very rare term.

To improve the operations of text processing systems, it has been suggested that the original document vocabulary be expanded by adding related or associated terms not originally present in the available text samples. Two main types of vocabulary relationships can be recognized in this connection, known respectively as <u>paradigmatic</u> and <u>syntagmatic</u> relations. [1] The paradigmetric relations cover term associations, such as synonyms and hierarchical inclusion, that always exist between particular terms regardless of the context in which these terms are used. For example, a paradigmatic relation exists between the name of a country (say, France) and the capital city (Paris). Syntagmatic relations, on the other hand, are relations which are not valid outside some specified context. For example, a cause-effect relation may be detected in certain circumstances between "poison" and "death".

*Department of Computer Science, Cornell University, Ithaca, NY 14853.
This study was supported in part by the National Science Foundation under grants IST 83-16166 and IST 85-44189.

The paradigmatic relations may be identified by using preconstructed dictionaries, or <u>thesauruses</u>, containing schedules or groupings of related terms or concepts. The syntagmatic relations, on the other hand, must be derived by analyzing particular text samples and extracting the term relationships specified in these texts.

Various methods are outlined in the next section for utilizing paradigmatic and syntagmatic term associations in text processing systems, and the effectiveness of the methods is assessed using available experimental output.

2. Associative Text Processing Methods

A) Thesaurus Operations

A <u>thesaurus</u> is a word grouping device which provides a hierarchical and/or a clustered arrangement of the vocabulary for certain subject areas. Thesauruses are used in text processing for three main purposes [2]:

a) as authority lists where the thesaurus normalizes the indexing vocabulary by distinguishing terms that are allowed as content identifiers from the remainder of the vocabulary;

b) as grouping devices where the vocabulary is broken down into classes of related, or synonymous terms, as in the traditional Roget's thesaurus;

c) as term hierarchies where more general terms are broken down into groups of narrower terms, that may themselves be broken down further into still narrower groups.

When a thesaurus is available for a particular subject area, each term found in a document can be used as an entry point into the thesaurus, and additional (synonymous or hierarchically related) terms included in the same thesaurus class as the original can be supplied. Such a thesaurus operation normalizes the vocabulary and provides additional opportunities for matches between query and document vocabularies. The vocabulary expansion tends to enhance the search recall (the proportion of relevant materials actually retrieved as a result of a search process).

When the subject area is narrowly circumscribed and knowledgeable subject experts are available, useful thesaurus arrangements can be manually constructed by human experts that may provide substantial enhancements in retrieval effectiveness. Table 1 shows the average search precision (the proportion of retrieved materials actually relevant) obtained at certain fixed recall points for a collection of 400 documents in engineering used with 17 search requests. In that case, the performance of a manually constructed thesaurus (the Harris Three thesaurus) is compared with a content analysis system based on weighted word stems extracted from document and query texts. The output of Table 1 shows that at the high recall end of performance range, the thesaurus provides much better retrieval output than the word stem process. [3]

Recall	Average Search Precision: Weighted Word Stems	Harris Three Thesaurus	
.1	.9563	.9735	+ 2%
.3	.7986	.8245	+ 3%
.5	.6371	.7146	+11%
.7	.4877	.6012	+19%
.9	.3426	.4973	+31%
			+13%

Sample Thesaurus Performance
(IRE Collection, 400 documents, 17 queries)
Table 1

While the use of thesauruses is widely advocated as a means for normalizing the vocabulary of document texts, no consensus exists about the best way of constructing a useful thesaurus. It was hoped early on, that thesaurses could be built automatically by studying the occurrence characteristics of the terms in the documents, and grouping into common thesaurus classes those terms that co-occur sufficiently often in the text of the documents: [4]

> "the statistical material that may be required in the manual compilation of dictionaries and thesauruses may be derived from the original texts in any desired form and degree of detail."

Later is was recognized that thesauruses constructed by using the occurrence characteristics of the vocabulary in the documents of a collection do not in fact provide generally valid paradigmatic term relations, but identify instead locally valid syntagmatic relations derivable from the particular document environment. [5] To utilize the conventional paradigmatic term relations existing in particular subject areas, the vocabulary arrangements must effectively be constructed by subject experts using largely ad-hoc procedures made up for each particular occasion. The thesaurus method is therefore not generally usable in operational environments.

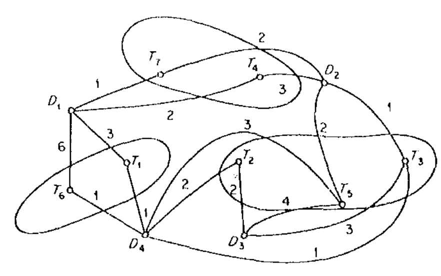
B) Automatic Term Associations

While generally valid thesauruses are difficult to build, locally valid <u>term association maps</u> can be generated automatically by making use of similarity measurements between pairs of terms based, for example, on the number of documents in which the terms co-occur in the documents of a collection. The number of common sentences in which a pair of words can be found may also be taken into account, as well as some measure of proximity between the words in the various texts. Using similarity measurements between word pairs, term association maps can be constructed, and these may be displayed and used by the search personnel to formulate useful query statements, and to obtain expanded document representations. [6,7]

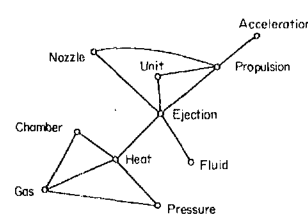
Terms assigned to documents

Document vectors	T_1	T_2	T_3	T_4	T_5	T_6	T_7 ....
D_1	3	0	0	2	0	6	1
D_2	0	0	1	3	2	0	2
D_3	0	2	3	0	4	0	0
D_4	1	2	1	0	3	1	0

(a) Term-Document Matrix Showing Frequency of Terms Assigned to Documents



(b) Term-Document Graph for Matrix of Fig. 1(a)



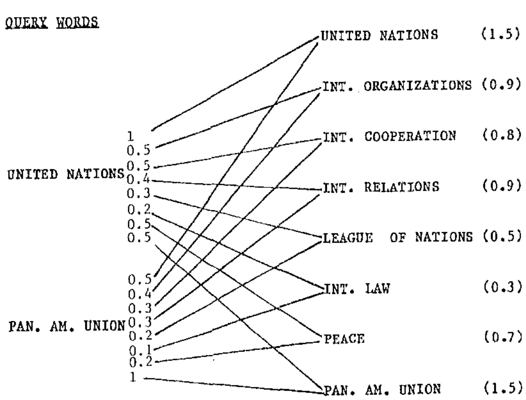
(c) Term Association Map

Term-Document Matrix and Graph and Corresponding Association Map (from [8] p.51)

Fig. 1

A sample assignment of terms to documents is shown in the matrix of Fig. 1(a). Fig. 1(b) shows the corresponding document-term graph where a line between term T_i and document D_j represents the corresponding term assignment appears in Fig. 1(b). Given the assignment of Figs. 1(a) and (b), term associations may be derived by grouping sets of terms appearing in similar contexts. For example, terms T_4 and T_7 may be grouped because these terms appear jointly in documents D_1 and D_2; similarity terms T_1 and T_6 appear in documents D_1 and D_4. The grouping operations may be used to obtain a term association map of the kind shown in Fig. 1(c), where associated terms are joined by lines on the map.

The operations of a typical associative retrieval system are illustrated in Fig. 2. [7] The original query words are listed on the left-hand side of Fig. 2, and the derived associated terms are shown on the right. The value of a given associated term--for example, "International Organizations"-- is computed as the sum of the term association values between the given term and all original query terms (0.5 for "United Nations" plus 0.4 for "Pan American Union" in the example of Fig. 2). Finally the retrieval value of a document is computed as the sum of the common term association values for all matching terms that are present in both queries and documents. Many variations are possible of the basic scheme illustrated in Fig. 2; in each case, the hope is that valid term associations would make it possible to achieve a greater degree of congruence between document representations and query formulations.



Document 1 :	UNITED NATIONS	1.5
	LEAGUE OF NATIONS	0.5
	PAN. AM. UNION	1.5
	INT. COOPERATION	0.8

		4.3
Document 2 :	UNITED NATIONS	1.5
	PAN. AM. UNION	1.5
	INT. LAW	0.3

		3.3

.
.
.

Associative Retrieval Example
(from Giuliano [6])

Fig. 2

In practice, it is found that the use of term associations can improve the search recall by providing new matches between the term assigned to queries and documents that were not available in the original query and document. In addition, the search precision can also be enhanced by reinforcing the strength of already existing term matches. [5] Unfortunately, the experimental evidence indicates that only about 20 percent of automatically derived associations between pairs of terms are semantically significant; the associative indexing process does not therefore provide guaranteed advantages in retrieval effectiveness.

Table 2 shows a typical evaluation output for a collection of 400 documents in engineering used with 17 search requests. The output of Table 2 shows that the automatic term associations provide an increase in average search precision only at the high recall end of the performance range. Overall, the average search precision decreases by 13 percent for the collection used in Table 2. [8, p. 130]

Recall	Average Search Precision: Weighted Word Stems	Automatic Term Associations	
.1	.9563	.7385	-23%
.3	.7986	.5844	-27%
.5	.6371	.5187	-19%
.7	.4877	.4452	- 9%
.9	.3426	.3794	+11%
			-13%

Sample Associative Indexing Performance
(IRE Collection, 400 documents, 17 queries)

Table 2

More recently other vocabulary expansion experiments have been conducted using associated terms derived by statistical term co-occurrence criteria. [9-10] Once again, the evaluation results were disappointing: [10]

> "Our results on query expansion using the NPL data are disappointing. We have not been able to achieve any significant improvements over nonexpansion. We have repeated previous experiments in which the query was expanded, and the resulting set of search terms then weighted... Once again the results have been conflicting..."

The conclusions derived from the available evidence indicate that the vocabulary expansion techniques which add to the existing content identifiers related terms specified in a thesaurus, or derived by term co-occurrence measurements, do not provide methods for improving retrieval effectiveness. Generally valid thesauruses for large subject areas are difficult to generate and the automatic term co-occurrence procedures do not offer adequate quality control. Efforts to enhance the recall performance of search systems must therefore be based on different techniques designed to generate indexing vocabularies of broader scope, including especially word stem generation and suffix truncation methods.

C) Term Phrase Generation

The vocabulary expansion methods described up to now are designed principally to improve search recall. Search precision may be enhanced by using narrow indexing vocabularies consisting largely of term phrases replacing the normally used single terms. Thus "computer science" or "computer programming" could replace a broader term such as "calculator" or "computer". The recognition and assignment of term phrases poses much the same problems as the previously described generation of term associations and the expansion of indexing vocabularies. In particular, an accurate determination of useful term phrases, and the rejection of extraneous phrases, must be based on syntactic analyses of query and document texts supplemented by semantic components valid for the subject areas under consideration. Unfortunately, complete linguistic analyses of topic areas of reasonable scope are unavailable for reasons of efficiency as well as effectiveness. In practice, it is then necessary to fall back on simpler phrase generation methods in which phrases are identified as sequences of co-occurring terms with appropriate statistical and/or syntactic properties. In such simple phrase generation environments quality control is, however, difficult to achieve.

The following phrase generation methods are of main interest:

a) statistical methods where each phrase head (the main phrase component) has a stated minimal occurrence frequency in the texts under consideration, and each phrase component exhibits another stated minimal occurrence frequency, and the distance in number of intervening words between phrase heads and phrase components is limited to a stated number of words;

b) a simple syntactic pattern matching method where a dictionary search method is used to assign syntactic indicators to the text elements, and phrases are then recognized as sequences of words exhibiting certain previously established patterns of syntactic indications (e.g. adjective-noun-noun, or preposition-adjective-noun); [11-12]

c) a more complete syntactic analysis method supplemented if possible by appropriate semantic restrictions to control the variety of permitted syntactic phrase constructions for the available texts. [13,14]

When statistical phrase generation methods are used, a large number of useful phrases can in fact be identified, together unfortunately with a large number of improper phrases that are difficult to reject on formal grounds. For example, given a query text such as

> "hemophilia and christmas disease, especially in regard to the specific complication of pseudotumor formation (occurrence, pathogenesis, treatment, prognosis)"

it is easy to produce correct phrase combinations such as "christmas disease" and "pseudotumor formation". At the same time the statistical phrase formation process produces inappropriate patterns such as "formation occurrence" and "complication formation". [15] Overall a statistical phrase formation process will be of questionable usefulness.

Table 3 shows a comparison of the average search precision results for certain fixed recall values between a standard single term indexing system, and a system where the single terms are supplemented by statistically determined phrase combinations. The output of Table 3 for four different document collections in computer science (CACM), documentation (CISI), medicine (MED) and aeronautics (CRAN) shows that the phrase process affords modest average improvements for three collections out of four. [15] However, the improvement is not guaranteed, and is in any case limited to a few percentage points in the average precision.

The evaluation results available for the syntax-based methods are not much more encouraging. [16] The basic syntactic analysis approach must be able to cope with ordinary word ambiguities (lamp base, army base, baseball base), the recognition of distinct syntactic constructs with identical meanings, discourse problems exceeding sentence boundaries such as pronoun referents from one sentence to another,

Recall	CACM 3204		CISI 1460		MED 1033		CRAN	
	Single Terms	Phrases	Single Terms	Phrases	Single Terms	Phrases	Single Terms	Phrases
.1	.5086	.5427	.4919	.4590	.8038	.7970	.7526	.7540
.3	.3672	.3971	.3118	.2999	.6742	.7064	.5184	.5385
.5	.2398	.2527	.2320	.2222	.5447	.5529	.3714	.3989
.7	.1462	.1462	.1504	.1283	.4082	.4166	.2301	.2431
.9	.0711	.0759	.0739	.0630	.2057	.2056	.1313	.1328
Average Improvement		+6.8%		-8.6%		+1.6%		+4.1%

Comparison of Single Term Indexing with Statistical Phrase Indexing for Four Sample Document Collections

Table 3

and the difficulties of interpreting many complex meaning units in ordinary texts. An illustration of the latter kind is furnished by the phrase "high frequency transistor oscillator", where it is important to avoid the interpretation "high frequency transistor" while admitting "transistor oscillator" and "high frequency oscillator". A sophisticated syntactic analysis system with substantial semantic components was unable in that case to reject the extraneous interpretations "frequency transistor oscillators which are high (tall)" and "frequency oscillators using high (tall) transistors". [17]

In addition to the problems inherent in the language analysis component of a phrase indexing system, a useful text processing component must also deal with phrase classification, that is the recognition of syntactically distinct patterns that are semantically identical ("computer programs", "instruction sets for computers", "programs for calculating machines"). The phrase classification problem itself raises complex problems that are not close to solution. [18]

In summary, the use of complex identifying units and term associations in automatic text processing environments is currently hampered by difficulties of a fundamental nature. The basic theories needed to construct useful term grouping schedules and thesauruses valid for particular subject areas are not sufficiently developed. As a result, the effectiveness of associative retrieval techniques based on term grouping and vocabulary expansion leaves something to be desired. The same is true of the syntactic and semantic language analysis theories used to generate a large proportion of the applicable complex content descriptions and phrases, and to reject the majority of extraneous term combinations.

The question arises whether any retrieval situations exists in which it is useful to go beyond the basic single term text analysis methodology, consisting of the extraction of single terms from natural language query and document texts. This question is examined in the remaining section of this note.

3. The Usefulness of Complex Text Processing

Three particular text processing situations can be identified where term association techniques have proved to be useful. The first one is the well-known relevance feedback process where initial search operations are conducted with preliminary query formulations obtained from the user population. Following the retrieval of certain stored text items, the user is asked to respond by furnishing relevance assessments for some of the previously retrieved items; these relevance assessments are then used by the system to construct new, improved query formulations which may furnish additional, hopefully improved, retrieval output. In particular, the query statements are altered by adding terms extracted from previously retrieved items that were identified as relevant to the user's purposes, while at the same time removing query terms included in previously retrieved items designated as nonrelevant.

The relevance feedback methodology represents an associative retrieval technique, since new query terms are obtained from certain designated documents that hopefully are related to the originally available formulations. Relevance feedback techniques have been used with vector queries formulated as sets of search terms [9, 19-20], and more recently with Boolean queries. [21] The effectiveness of the feedback procedure has never been questioned.

Table 4 shows typical evaluation output for four different document collections in terms of average search precision at ten recall points (from a recall of 0.1 to a recall of 1.0 in steps of 0.1) averaged over the stated number of user queries. The output of Table 4 applies to Boolean queries with binary weighted terms. [21] The improvements in retrieval precision due to the user feedback process ranges from 22% to 110% for a single search iteration. When the feedback process is repeated three times, the improvement in search precision increases to 63% to 207%. Evidently, the user relevance information which applies to particular queries at particular times makes it possible to find a sufficient number of interesting term associations to substantially improve the retrieval output.

A second possibility for generating improved retrieval output consists in limiting the analysis effort to the user query formulations instead of the document texts. In a recent study, term phrases were first extracted from natural language query texts using a simple, manually controlled, syntactic analysis process. These query phrases were then recognized in document texts by a rough pattern

	Medlars 1033 30 queries	CISI 1460 35 queries	CACM 4204 52 queries	Inspec 12684 77 queries
Original Boolean Queries	0.2065	0.1118	0.1798	0.1159
First Iteration Relevance Feedback	0.4322 (+110%)	0.1367 (+22%)	0.2550 (+42%)	0.1522 (+31%)
Third Iteration Relevance Feedback	0.6334 (+207%)	0.1827 (+63%)	0.3217 (+79%)	0.1933 (+67%)

Average Search Precision at 10 Recall Points for One Iteration and Three Iterations of Relevance Feedback (4 document collections)

Table 4

matching procedure distinguishing pairs and triples of terms occurring in the same phrases of documents, and pairs and triples of terms occurring in the same sentences of documents. [22] Whenever a phrase match is obtained between a query phrase and a document text, the retrieval weight of the document is appropriately increased.

An evaluation of such a manually controlled syntactic phrase recognition system based on query statement analysis reveals that substantial improvements in retrieval effectiveness are obtainable for the phrase assignments, compared with the single term alternatives. Table 5 shows average search precision values at five recall levels for 25 user queries used with the CACM collection in computer science. [22] On average the query analysis system raises the search precision by 32 percent.

Recall	Average Search Precision		
	Weighted Single Terms	Weighted Single Terms and Phrases	
0.1	0.555	0.625	+13%
0.3	0.271	0.355	+31%
0.5	0.211	0.265	+26%
0.7	0.064	0.085	+33%
0.9	0.038	0.060	+58%
			+32%

Average Search Precision for
Query Statement Phrase Analysis
(CACM Collection, 25 Queries)

Table 5

The special processing described up to now is user related in the sense that user query formulations and user relevance assessments are utilized to improve the retrieval procedures. The last possibility for the use of complex information descriptions consists in incorporating stored knowledge representations covering particular subject areas to enhance the descriptions of document and query content. [23-25] Various theories of knowledge representation are current, including for example, models based on the use of frames representing events and descriptions of interest in a given subject.

Frames designating particular entities may be represented by tabular structures, with open "slots" filled with attributes of the entities, or values of attributes. Relationships between frames are expressed by using attributed that are themselves represented by other frames, and by adding links between frames. Frame operations can also be introduced to manipulate the knowledge structure when new facts or entities become known, or when changes occur in item relationships. There is some evidence that when the knowledge base needed to analyze the available texts is narrowly circumscribed and limited in scope, useful frame structures can in fact be intellectually prepared to enhance the retrieval operations. [26]

However, when the needed topic area is not of strictly limited scope, the construction of useful knowledge bases is much less straightforward and the knowledge-based processing techniques become of limited effectiveness. It has been suggested that in these circumstances, the system user himself might help in building the knowledge structures. [27] While this remains a possibility, it is hard to imagine that untrained users can lay out the subject knowledge of interest in particular areas and specify concept relationships such as synonyms, generalizations, instantiations, and cross-references with sufficient accuracy. In any case, no examples exist at the present time where user constructed knowledge bases have proved generally valid for different collections in particular subject areas. In fact, the situation appears much the same as it was thirty years ago: it seems quite easy to build locally valid term association systems by ad-hoc means; these tools fail however in somewhat different environments, and do not furnish reliable means for improving text processing systems in general.

For the foreseeable future, text processing systems using complex information identifications and term associations must therefore be limited to narrowly restricted topic areas, or must alternatively be based on simple user inputs, such as document relevance data, that can be furnished by untrained users without undue hardship.

References

[1] J.C. Gardin, Syntol, in Systems for the Intellectual Organization of Information, S. Artandi, editor, Vol. 2, Rutgers University, New Brunswick, NJ, 1965.

[2] M.E. Stevens, Automatic Indexing: A State of the Art Report, NBS Monograph 91, National Bureau of Standards, Washington, DC, 1965.

[3] G. Salton and M.E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, 15:1, January 1968, 8-36.

[4] H.P Luhn, Auto-Encoding of Documents for Information Retrieval Systems", M. Boaz, editor, Modern Trends in Documentation, 1959, 45-58.

[5] M.E. Lesk, Word-Word Associations in Document Retrieval Systems, American Documentation, 20:1, January 1969, 27-38.

[6] L.B. Doyle, Semantic Road Maps for Literature Searchers, Journal of the ACM, 8, 1961, 553-578.

[7] V.E. Giuliano, Automatic Message Retrieval by Associative Techniques, in Joint Man-Computer Languages, Mitre Corporation Report SS-10, Bedford, MA, 1962, 1-44.

[8] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill Book Company, New York, 1968.

[9] D.J. Harper and C.J van Rijsbergen, An Evaluation of Feedback in Document Retrieval using Cooccurrence Data, Jl. of Documentation, 34:3, September 1978, 189-216.

[10] S.E. Robertson, C.J. van Rijsbergen, and M.F. Porter, Probabilistic Models of Indexing and Searching, in Information Retrieval Research, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, and P.W. Williams, editors, Butterworths, London, 1981, 35-56.

[11] M. Dillon and A.S. Gray, FASIT: A Fully Automatic Syntactically Based Indexing System, Journal of the ASIS, 34, 1983, 99-108.

[12] G. Salton, Automatic Phrase Matching, in Readings in Automatic Language Processing, D.G. Hays, editor, Am. Elsevier Publishing Co., New York, 1966.

[13] R. Grishman, Natural Language Processing, Journal of the ASIS, 35, 1984, 291-296.

[14] P.J. Hayes and J.G. Carbonell, A Tutorial on Techniques and Applications for Natural Language Processing, Technical Report CMU-CS-83-158, Carnegie-Mellon University, Pittsburgh, PA, 1983.

[15] J.L. Fagan, Automatic Phrase Indexing for Text Passage Retrieval and Printed Subject Indexes, Technical Report, Department of Computer Science, Cornell University, Ithaca, NY, May 1985.

[16] G. Salton, The Smart Retrieval System - Experiments in Automatic Document Processing, G. Salton, editor, Prentice Hall Inc., Englewood, Cliffs, NJ, 1971, 207-208.

[17] K. Sparck Jones and J.I. Tait, Automatic Search Term Variant Generation, Journal of Documentation, 40:1, March 1984, 50-66.

[18] C.D. Paice and V. Aragon-Ramirez, The Calculation of Similarities between Multi-word Strings using a Thesaurus, Proc. RIAO-85 Conference, Grenoble, France, 1985, 293-319.

[19] J.J. Rocchio, Jr., Relevance Feedback in Information Retrieval, in The Smart System - Experiments in Automatic Document Processing, Prentice Hall, Inc., Englewood Cliffs, NJ, 1971, Chapter 14.

[20] E. Ide, New Experiments in Relevance Feedback, in The Smart System - Experiments in Automatic Document Processing, G. Salton, editor, Prentice Hall Inc., Englewood Cliffs, NJ, 1971, Chapter 16.

[21] G. Salton, E.A. Fox and E. Voorhees, Advanced Feedback Methods in Information Retrieval, Journal of the ASIS, 36:3, 1985, 200-210.

[22] A.F. Smeaton, Incorporating Syntactic Information into a Document Retrieval Strategy: An Investigation, Technical Report, Department of Computer Science, University College, Dublin, Ireland, 1986.

[23] M. Minsky, A Framework for Representing Knowledge, P.H. Winston, editor, The Psychology of Computer Vision, McGraw Hill Book Co., NY, 1975, 211-277.

[24] R.C. Schank and R.P. Abelson, Scripts, Plans, Goals and Understanding, Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

[25] R.J. Brachman and B.C. Smith, Special Issue on Knowledge Representation, SIGART Newsletter, No. 70, February 1980.

[26] M.K. diBenigno, G.R. Cross and C.G. deBessonet, COREL - A Conceptual Retrieval System, Technical Report, Louisiana State University, Baton Rouge, LA, 1986.

[27] W.B. Croft, User Specified Domain Knowledge for Document Retrieval, Technical Report, Computer Science Department, University of Massachusetts, Amherst, MA, 1986.