The Use of Standardized Documentary Data in Automatic Information Dissemination

G. Salton

Technical Report
No. 68-12
March 1968

Department of Computer Science
Cornell University, Ithaca, N. Y.

The Use of Standardized Documentary Data in Automatic Information Dissemination

G. Salton*

Abstract

It is likely that future operating information retrieval systems may be based on automatic information analysis methods instead of manual indexing, and on search procedures which allow the user to interact with the system during the search process. The effectiveness of the required analysis and search operations depends to some extent on the availability in machine readable form of standardized information concerning the make-up and content of each stored document.

An author-prepared standard manuscript documentation unit, furnished with each manuscript, may simplify the information retrieval and dissemination operations, and improve their effectiveness. The design of such a documentation unit is covered and its use is explained for indexing, classification, vocabulary normalization, searching, and retrieval.

1. Introduction

Over the last few years, the design and operations of large-scale information systems has become of concern to a considerable segment of the scientific and technically oriented population. Furthermore, as the amount and complexity of the available information has continued to grow, the use of mechanized or partly-mechanized procedures for various information storage and retrieval tasks has also become more widespread. As a result,

* Department of Computer Science, Cornell University, Ithaca, New York.

This study was prepared for submission to the IEEE Transactions on Engineering Writing and Speech, and was supported in part by the National Science Foundation under grant GN-495.

several large information systems are now in operation in which the information search operations, consisting normally in a comparison of incoming search requests with stored documents, is carried out automatically. Typioal examples in the United States are the NASA Scientific and Technical Information Facility, and the MEDLARS system at the National Library of Medicine.

While it is thus operationally possible to search rapidly through vast storage files, often containing many hundreds of thousands of items, most of the operations other than the search itself are normally performed manually with the help of human experts. In particular, all the content analysis and indexing operations, leading to the assignment of suitably chosen combinations of index terms to the stored documents and to incoming search requests are in general performed by specialists who know the given subject area as well as the performance characteristics of the retrieval environment within which they operate.

The extensive systems evaluation work which has been carried out over the last few years, under operational and laboratory conditions leads to two main conclusions: first, that the existing partly-manual systems are remarkably successful in isolating from the large mass of stored material many of the items which eventually prove pertinent to the requestor's information needs; and, second, that the presently achievable performance is far short of what is theoretically desirable.

The search and retrieval failures which are found to occur in many of the presently operating systems (that is, failures to retrieve material which is relevant to the users' needs, or failures to reject material which is not relevant) are due to many diverse causes. By far, the largest number of failures appear, however, to be caused in one way or another by inadequacies in the information indexing and analysis system. Retrieval errors can thus occur as a result of basic insufficiencies in the indexing vocabulary — for example, the fact that appropriate terms are not available in the vocabulary to properly express the content of a document or search request. Alternatively, the analysis process itself may be at fault, as happens when the indexing terms assigned to an information item are not sufficiently specific, or do not reflect all aspects of the information content (lack of exhaustivity). Finally, many errors are indirectly caused by a lack of interaction between the user and the information officer charged with the formulation of the user's search request [1,2,3,4].

It is not likely that any one simple change in systems design will be sufficient by itself to remove all retrieval errors. However, a number of different approaches are nevertheless available which may lead to improvements in systems performance. The first consists in replacing at least some of the analysis and indexing operations which are now conducted manually by alternative automatic and semi-automatic methods. The second relates to the generation of more powerful thesauruses and dictionaries to be used for language normalization. The third would replace the present search operations, conducted in the absence of the user, by user-controlled semi-automatic search methods in which the customer himself helps during the search by furnishing appropriately refined reformulations of the queries. The last approach toward the possible redesign of information retrieval systems consists in bringing into the system only those documents and information items which have already been standardized to some extent in appearance and content.

The introduction with each document of a standard manuscript documentation unit, prepared by the author of each manuscript, is a response to this last requirement. However, as will be seen, the documentation units can also be used as an aid in the generation of better thesauruses, and as a guide in the analysis and indexing process.

In the present study, a number of problems are first examined which arose in automatic text processing, including in particular, the automatic input problem, the question of thesaurus construction, and certain aspects of language and content analysis. Some indications are given of the effectiveness of various types of text processing methods, and conclusions are drawn concerning the preferred form of texts for storage and retrieval purposes. This leads to the design of a manuscript document unit to be made available with each information item, when a document is first produced.

The design of the documentation unit is described and various applications are examined for such units. Some preliminary experience is also given concerning the use of documentation units for vocabulary standardization and dictionary construction.

2. Automatic Text Processing

A) Input Conversion

At the present time, several manual typing or keying operations are required during the various processing stages of a manuscript: first by the author during the initial steps, and again following the review process if corrections are needed, and again during typesetting, and once more for inclusion into various secondary dissemination media, such as index volumes, catalogs, abstract journals, etc., and again for introduction into a storage and retrieval system.

Ideally, it would be useful, if each text were handled fully automatically, or failing this, if a single typing operation were to be sufficient for all subsequent processes. There exist automatic print readers which read text by using optical or magnetic character recognition techniques. However, commercially available, reliable reading equipment is still largely restricted to specially identified typefonts and to characters obeying certain standards of quality and appearance. In particular, most readers cannot deal with large character sets, multiple typefonts, skewed type, variable sizes, and variable printing densities of the kind currently occurring in many printed texts. The alternative of using voice input is even further removed from operational practice, because of the difficulties in automatically recognizing ordinary speech.

The alternative arrangement consists in using a single keying operation to convert a text into machine-readable form, thereby producing for example a paper tape or a set of punched cards. Following this, all subsequent operations are then conducted starting from the available tape or cards. Automatic editing systems can be used, in particular, which generate from the original input a new altered tape, including format alterations and corrections. This can be followed by automatic typesetting operations, where the input tape is first converted, under computer control, into a new format including the special symbols and instructions necessary to control a typesetting machine, and this intermediate tape is then used to drive an automatic typesetter [5,6,7].

Subsequent operations such as the production of indexes, abstracts and citations, as well as the assignment of content identifiers can then be initiated using an intermediate output tape produced as a result of some prior operation. During these analysis tasks, it is normally essential to introduce methods for vocabulary normalization, since items dealing with similar subject matter should be assigned similar content identifiers, and should be transformed into similar reduced forms.

The dictionary construction problem which results is treated in the next subsection.

B) Dictionary Construction

To insure a reasonably consistent assignment of content identifiers, authority lists of various kinds may prove useful, including negative dictionaries containing lists of terms which should <u>not</u> be used for indexing purposes, thesauruses and synonym dictionaries which group certain synonymous or related terms under common group headings, phrase dictionaries which include strings of words or phrases useful for content identification, and hierarchical arrangements of terms. A typical thesaurus arrangement used in the automatic SMART document retrieval system is shown in Table 1 [8,9,10]. Normally, each word occurring in a text is looked up in the thesaurus, and replaced by one or more "concept numbers" reflecting its meaning. Concept numbers above 32,000 are used in Table 1 to identify terms which are not suitable as content identifiers. The syntax codes in column 3 of Table 1(b) are included for syntactic analysis purposes.

The problem of dictionary construction is one of major proportions, not only because of the relative magnitude of the task, but also because most dictionaries which might be built by random methods would not be very useful for retrieval. The previously mentioned system evaluation studies [2,3] apparently lead to the conclusion that the standard accepted way of generating authority lists — by a committee of experts who decide on the inclusion and grouping of terms — may not produce the most effective results. Often such committees have no clear idea of the ultimate use of the dictionary, and the product obtained reflects compromises made in order to come to an agreement regarding form and content of the listing.

A possible improvement over the standard manual dictionary construction procedures may be provided by methods using information derived from sample document collections in the subject field under consideration. In particular, word frequency lists can be generated automatically, as well as concordances showing the word occurrences in their context. Thesaurus entries may then be chosen from these lists, and groupings of related words may be constructed either manually or automatically. The thesaurus construction principles used for this purpose with the SMART system are summarized in Table 2. As will be seen later, fully-automatic methods can also be used to generate term associations, and terms with a sufficiently high degree of association can then be grouped into a common thesaurus class [11].

Once a thesaurus is available, a great many automatic, or semi-automatic, procedures can be used which utilize the stored dictionary for purposes of language analysis [12,13]. These questions are further discussed in the next few sections.

C) Language Analysis

In a document processing context, the language analysis problem generally consists in the generation of a set of content identifiers — sometimes called keywords or index terms — designed to reflect document content. If a set of keywords is assigned to each stored document and to each search request, the information search is simple reducéd to a comparision between keyword sets assigned to documents and to queries, followed by the retrieval of those documents whose keywords are sufficiently similar to the query keywords.

Normally, the keywords are manually chosen by trained indexers, or subject experts, using for this purpose language normalization tools such as those described in the previous section. Some experimental text processing systems have, however, been designed, in which an attempt is made to replace the intellectual indexing effort used in the conventional situation by a fully-automatic computer analysis of the document and query texts. Specifically, a variety of language analysis procedures — such as suffix cut-off methods, thesaurus look-up, phrase generation methods, statistical term associations, syntactic analysis, and others — are used to reduce document and query texts into analyzed concept (or term) vectors. The concept vectors attached to the documents are then matched with the vectors derived from the search requests, and documents, arranged in decreasing query correlation order, are submitted to the user as answers to the query.

In addition to providing an experimental retrieval facility, some of these experimental systems also make it possible to evaluate the retrieval effectiveness of many of the automatic language analysis systems used to generate the term vectors, and to compare the usefulness of the mechanically generated concepts with those manually assigned by the subject experts. This can be done by processing the same search requests against the same document collections, and comparing the results obtained when different analysis procedures are used. In the SMART system, such an evaluation is based on two measures of retrieval effectiveness, known as recall and precision, respectively, where recall corresponds to the proportion of relevant material actually retrieved, while precision is the amount of retrieved material actually relevant. By varying the number of items retrieved in response to a search request, several recall-precision pairs are obtained for each query. These can be averaged over many search requests, and recall can be plotted on a graph against precision to produce recall-precision curves such as those shown in Fig. 1.

In an ideal system, all relevant materials would be retrieved while at the same time all nonrelevant items would be rejected. Under those circumstances, the normal recall-precision graph shrinks to a single point, where both recall and precision are equal to 1. In practice, the performance is generally far away from the ideal; in fact, the usual operating situation is one where more relevant items can be retrieved only at the cost of also retrieving more nonrelevant ones. Thus, as the recall goes up, the precision goes down, and the recall-precision curves take the monotonically decreasing aspect illustrated in Fig. 1.

Three different text processing systems are compared in Fig. 1, by showing the average performance for 34 queries processed by SMART with a collection of 780 documents in computer engineering [2]. The graph of Fig. 1(a) compares a word stem matching process — where weighted word stems contained in document abstracts are matched against word stems occurring in the search requests — against a thesaurus process, where the word stems are first normalized by reference to a thesaurus prior to the comparison between queries and documents. It may be seen that the thesaurus curve is closer to the ideal performance range in the upper right hand corner of the graph, thus demonstrating that, on the average, the use of a thesaurus does in fact improve retrieval performance.

The small table below the recall-precision graph contains the average precision values obtained at several recall levels, and the rightmost column of each table represents the statistical significance of the differences between the corresponding measures for the methods being compared. Specifically, the value in the rightmost column is the probability, computed by a standard sign test, that the respective measures for the two methods (for example the average precision value at recall of 0.5) might have originated from the same distribution. A probability of 0.05 or less is normally accepted as a reliable indication that the assumption of equivalence can be rejected, since the probability is then sufficiently small that differences as large as the ones actually noted might have been produced by random variations of the input data. For the case of the stem-thesaurus comparison, the thesaurus provides significantly better performance over most of the recall-precision range.

The middle graph of Fig. 1(b) compares two word stem matching procedures, the first one using document abstracts as input, and the other using only document titles. It is seen that the titles are not, on the average, very effective for retrieval purposes, and that the performance obtained by using abstracts as a starting point for the analysis process is significantly better (except at the very low recall end of the curve).

The output of Fig. 1(c) compares an analysis procedure using a standard thesaurus with one including also phrases, rather than single terms, as information identifiers. The phrase process is based on the realization that many terms may be ambiguous when used in isolation, such as, for example, "computer" or "control", whereas the same terms in combination are more specific, as in "computer control". For the process represented in Fig. 1(c), all allowable phrases are entered in a preconstructed dictionary, and a phrase in recognized whenever all the component concepts are contained within the same sentence of a document. It is seen that the phrase assignment affords an improvement over the standard thesaurus method, although the increase in performance is not statistically significant for the 34 queries used in Table 1.

One might expect that an automatic text analysis of the type used in the SMART system would necessarily produce retrieval results which are much inferior to those obtained in a system based on manual indexing. In actual fact, the automatic environment makes it possible to use for analysis purposes relatively large sections of text, such as abstracts or summaries, thus insuring a high degree of indexing exhaustivity; furthermore, the importance of some assigned terms can be enhanced by automatic weighting methods, leading to a more sophisticated matching process between analyzed queries and documents than is normally possible. As a result, the benefits of the index language control supplied in the more conventional retrieval situation by the human indexers appear to be balanced by a deeper and more complex type of analysis available in an automatic environment; this is reflected by evaluation results which indicate that the search effectiveness of the fully-automatic systems is not inferior to that obtainable at present in a partly-manual system.

Consider, as an example, the output of Fig. 2, where an automatic word stem procedure is compared with a standard keyword matching process (termed "index stem" in Fig. 2). Average results are shown for 42 queries processed with a collection of 200 document abstracts in aerodynamics. For the output shown in Fig. 2, the manual indexing was performed by trained indexers as part of the Aslib-Cranfield research project [1]. It is seen in the figure that the keyword matching process does in fact perform slightly better over most of the performance range than the somewhat rudimentary automatic word stem matching procedure. The output of Fig. 2(b) demonstrates, however, that the differences in performance are not statistically significant, thus indicating that the effectiveness of two systems is comparable.

The analysis procedures illustrated in the output of Figs. 1 and 2 appear to lead to the following conclusions:

a) Since complete document abstracts are more effective for content analysis purposes than document titles alone, provisions should be made for generating standardized document abstracts, and using them as part of a text processing system;

b) Since a content analysis process based on a stored thesaurus, or synonym dictionary, is more effective than one based directly on the comparison of word stems occurring in a text, provision should be made for the construction of thesauruses for the various subject fields, and for their use in information processing.

c) Since phrases can be used advantageously as content identifiers, a phrase generation process should be implemented, and used in conjunction with other analysis methods.

A manuscript documentation unit whose properties reflect these requirements is introduced in the next section.

3. Manuscript Documentation Unit

A manuscript documentation unit (MDU) is a body of data comprising descriptive information about a given manuscript or document. Ideally, the unit should be available when a manuscript is first generated, and it should be used during the various stages of document processing. As such, the necessary information is best generated by the author of each document. Alternatively, the documentation unit might be generated manually by subject experts, or automatic computer-based procedures might be used in its construction. The documentation unit used for manuscripts submitted to ACM serial publications (ACM Journal and ACM Communications) consists of five parts:

a) A descriptive title;

b) author names and addresses;

c) an informative abstract;

d) content indicators of two types, including key words and key phrases, and category numbers taken from a classification schedule for the computer field;

e) citations to the relevant literature.

A draft of the instructions for the preparation of the MDU is shown in Fig. 3 and certain additional details concerning the documentation unit are included in Table 3.

On the whole, the instructions reflect the requirements imposed by the document analysis process described in the previous section. The following features may be of particular interest:

a) The <u>title</u> should reflect the document content, and should contain as many highly specific terms as possible; this obviously implies that cute or clever titles should not be used, since they hide rather than promote the real document content, and are thus ineffective in a retrieval environment, since titles are often used as indexing tools — for example, in a permuted index such as KWIC — the inclusion in a title of special symbols and formulas should be avoided;

b) <u>abstracts</u> may be used as principal inputs for content analysis purposes, and in many cases, they replace a short version of the manuscript; the abstract must therefore be informative, rather than indicative, and must include direct information concerning the object of the manuscript, the procedures used, and the results obtained;

c) the <u>classification (category) identifiers</u> are used as an initial tool for the grouping of documents into subject categories; a classified collection can be stored efficiently, and searched rapidly, assuming that only the most likely classes of documents are compared with each search request rather than the whole collection; for the ACM literature, the category numbers correspond to the subject classification used in identifying review articles published in Computing Reviews.

d) the instructions covering the choice and form of <u>key words and phrases</u> are patterned after the requirements arising from the phrase matching procedures that might be incorporated in an automatic retrieval system; thus, nouns should be used, as well as noun-noun and noun-adjective combinations, since the resulting word string combinations can almost always be matched with proper English phrases; on the other hand, adverbs, prepositions, conjunctions, and similar function words should not be used in phrase construction, because such particles have too many different uses, and a determination of what word combinations do, in fact, constitute proper phrases then becomes unnecessarily difficult; since many analysis systems cannot recognize negative descriptors, identifying subject areas <u>not</u> handled in the document, only "positive" key words should be used; furthermore, broad catch-all terms which normally match with all sorts of extraneous material (for example, the word "computer" in a collection in computer science) should not be assigned as content identifiers.

Several different types of auxiliary aids can be used to help in choosing appropriate entries for the documentation unit. In particular, glossaries and standard subject indexes can be consulted, as well as other documents known to be related. A check can also be made to see whether terms included in titles and abstracts may be suitable as key words or phrases. Finally, if no thesaurus is available for synonym recognition and language normalization, each term or phrase actually included in the key word set can be supplemented by synonymous and related constructions.

Assuming that the documentation unit so constructed actually reflects the normal criteria of accuracy and completeness, the unit can then be used to replace a great deal of complicated analysis, or alternatively, the results of a standard content analysis can be supplemented by the information contained in the documentation unit.

The following principal uses suggest themselves for the various components of the documentation unit:

a) document indexing: the keywords and phrases can be used directly for document indexing, either by means of a thesaurus with or without vocabulary control; each set of terms then functions as a "profile" or term vector representing the content of the corresponding document, and the term vector may be used directly for retrieval purposes in standard demand-search systems, as well as in systems for selective dissemination of information;

b) document classification: by matching each document vector against each other it is possible to construct groups or classes of related documents consisting of those documents whose corresponding term vectors are sufficiently similar; such a classification of items into groups is useful for the production of catalogs and bibliographic lists of various kinds; in addition, the classification may also serve as a basis for fast search strategies in which incoming search requests are matched against only those documents which are located in classes likely to be of interest [14,15].

c) thesaurus construction: associations of terms or key words can be defined based on the joint assignment of two or more terms to the documents in a collection, or on the joint assignment of terms to documents which have previously been grouped into a common subject category or class; once a measure of association is available for each pair of terms, it becomes possible to form term or thesaurus classes by grouping all terms with a sufficiently high degree of association; there is reason to believe that such a process provides a viable starting procedure for the construction of a classed thesaurus, and is also usable to expand an already existing thesaurus;

d) document retrieval: the category identifiers and key words and phrases are usable for retrieval purposes, by basing the search strategy on the retrieval of those documents whose term vectors are sufficiently close to the query; the terms originally assigned can be used directly for vector matching, or the available terms might be altered in various ways by thesaurus look-up, or by addition of related terms prior to the matching operation; the retrieval procedure may also be carried out with or without phrase identifiers;

e) interactive retrieval strategies: in many information systems planned for the future,facilities are provided for executing user-controlled search strategies in which the user himself, sitting at a special input-output device connected to the central information files, directs the search; this can be accomplished by displaying, for the user's attention, certain thesaurus excerpts that may be of help in rephrasing search requests which do not initially produce satisfactory results; alternatively, the user may be permitted to describe his information needs in greater detail, or to furnish qualitative judgments about the results of previous searches; stored dictionaries and documents grouped into classes are advantageous in approaching the desired subject areas rapidly [16,17].

f) secondary publications: all parts of the documentation unit are also useful for the production of secondary different classes. The category assignment is given in Fig. 4(b) and the corresponding document network in Fig. 4(a). Each branch in the graph of Fig. 4(a) represents the joint assignment of a given category identifier to the two documents whose numbers identify the nodes of the graph.

The key words and phrases which occur in at least two of the six documents assigned to category 4.12 (compilers and generators) are listed in Table 4. The listing indicates that a relatively large and specialized vocabulary can be accumulated from only a small sample of documents. As more documents are available, the conditions for vocabulary grouping can be tightened by requiring, for example, a larger number of joint key word assignments or a tighter document cluster.

Some of the problems arising in vocabulary construction are illustrated in Table 5. The upper part of the table covers a case where the breadth of the subject categories is too great to permit a profitable comparison of the vocabulary. Specifically, while categories 4.41 and 6.35 are common to both documents 24 and 29, the only matching key word is "communication". Obviously, a category such as 6.35 (input-output equipment) covers a multitude of different devices which may be relatively unrelated.

The lower part of Table 5 contains an example where the lack of joint vocabulary is due in part to a less than ideal key word assignment. In particular, only the term "computer" matches for documents 17 and 23 although they are jointly assigned to category 1.52 (university courses). It is clear that the author of document 17 has not followed the instructions concerning phrase generation.

A manuscript documentation unit has been introduced, and the help of authors has been enlisted in its construction with the thought that a number of small, relatively simple steps may produce a major impact in the information dissemination field, if used consistently and with imagination. It is hoped that this expectation may be confirmed after additional experience is obtained, and more usage will have been made of the proposed manuscript documentation unit.

References

[1] C. W. Cleverdon and E. M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 2 — Test Results, Aslib-Cranfield Research Project, Cranfield, 1966.

[2] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 15, No. 1, January 1968.

[3] F. W. Lancaster, Evaluation of the Operating Efficiency of MEDLARS, Final Report, National Library of Medicine, January 1968.

[4] M. M. Henderson, Evaluation of Information Systems: A Selected Bibliography with Informative Abstracts, NBS Technical Note 297, National Bureau of Standards, December 1967.

[5] L. F. Buckland, Machine Recording of Textual Information during the Publication of Scientific Journals, Report to the National Science Foundation, Inforonics Inc., May 1965.

[6] M. P. Barnett and K. L. Kelley, Computer Editing of Verbal Texts — The ES-1 System, American Documentation, Vol. 14, No. 2, April 1963.

[7] M. P. Barnett, Computer Typesetting — Experiments and Prospects, MIT Press, 1965.

[8] G. Salton and M. E. Lesk, The SMART Automatic Document Retrieval System — An Illustration, Communications of the ACM, Vol. 8, No. 6, June 1965.

[9] M. E. Lesk, Operating Instructions for the SMART Text Processing and Document Retrieval System, Scientific Report No. ISR-11 to the National Science Foundation, Section II, Department of Computer Science, Cornell University, Ithaca, N. Y., June 1966.

[10] G. Salton et al., Scientific Reports on the SMART System, Nos. ISR-11, ISR-12, ISR-13, Department of Computer Science, Cornell University, June 1966, June 1967, and January 1968.

[11] G. Salton, Information Dissemination and Automatic Information Systems, Proc. IEEE, Vol. 54, No. 12, December 1966.

References
(contd)

[12] P. G. Chapin, L. N. Gross, L. M. Norton, R. J. Beller, and C. T. Browne, SAFARI: An On-Line Text Processing System — User's Manual, Report MTP-60, Mitre Corporation, March 1967.

[13] M. Masterman, Man-aided Computer Translation from English into French using an On-Line System to Manipulate a Bilingual Conceptual Dictionary or Thesaurus, Conférénce Internationale sur le Traitement Automatique des Langues, Grenoble, August 1967.

[14] G. Salton, Search Strategy and the Optimization of Retrieval Effectiveness, FID-IFIP Conference on Mechanized Information Storage, Retrieval, and Dissemination, Rome, June 1967.

[15] J. J. Rocchio, Jr., Document Retrieval Systems — Optimization and Evaluation, Harvard University Doctoral Thesis, Scientific Report No. ISR-10 to the National Science Foundation, Harvard Computation Laboratory, March 1966.

[16] E. Ide, User Interaction with an Automated Information Retrieval System, Scientific Report No. ISR-12 to the National Science Foundation, Section VIII, Department of Computer Science, Cornell University, June 1967.

[17] G. Salton, Search and Retrieval Experiments in Real-Time Information Retrieval, Proceedings of IFIP Congress '68, Edinburgh, August 1968.

408 DISLOCATION JUNCTION MINORITY-CARRIER N-P-N P-N-P POINT-CONTACT RECOMBINE TRANSITION UNIJUNCTION	411 COERCIVE DEMAGNETIZE FLUX-LEAKAGE HYSTERESIS INDUCT INSENSITIVE MAGNETORESISTANCE SQUARE-LOOP THRESHOLD
409 BLAST-COOLED HEAT-FLOW HEAT-TRANSFER	412 LONGITUDINAL TRANSVERSE

a) Thesaurus Excerpt (Concept Class Order)

Text Words	Concept Numbers	Syntax Codes
BLOCK	663	070043040
BLUEPRINT	58	070043
BOMARC	324	070
BOMBARD	424 0343	043
BOMBER	346	070
BOND	105	070043
BOOKKEEPING	34	070
BOOLEAN	20	001
BORROW	28	043
BOTH	32178	008080012
BOUND	523 0105	070043134135
BOUNDARY	524	070
BRAIN	404 0235	070
BRANCH	48 0042	070042
BRANCHPOINT	23	070
BREAK	380	043040070
BREAKDOWN	689	070
BREAKPOINT	23	070
BRIDGE	105 0458 0048	070043
BRIEF	32232	001043071
BRITISH	437	001071
BROAD-BAND	312	001071

b) Thesaurus Excerpt in Alphabetic Order

Sample Thesaurus Format
Table 1

Type of Term	Thesaurus Rule
Very rare terms	do not place into separate categories in the thesaurus, but combine if possible with other rare terms to form larger classes (low frequency categories provide few matches between stored items and search requests)
Very common terms	high-frequency terms should be either eliminated since they provide little discrimination, or should be placed into synonym classes of their own so as not to submerge other terms with which they might be grouped
Terms of no technical significance	terms which have no special significance in a given technical area (such as "begin", "automatic", "system", etc. in the computer science area) should be excluded from the thesaurus
Ambiguous terms	ambiguous terms should be entered into the thesaurus only in those senses likely to occur in the given subject area

Sample Thesaurus Construction Rules

Table 2

Documentation Unit	Instructions
Descriptive Title	6 to 12 words; use highly specific terms; avoid special symbols and formulas; avoid cute or clever titles;
Informative Abstract	150 to 200 words; short, direct and complete sentences state object of work, summarize results; give principal conclusions; state whether focus is practical or theoretical, and whether review, survey or tutorial;
Category Numbers	use as many group numbers from classification schedule as applicable; specify interpretation for each category;
Keywords and Phrases Properties	use nouns, or noun-noun and noun-adjective combinations; use up to three terms per phrase; avoid prepositions and hyphens; avoid broad catchall terms; avoid negative terms describing what is not done in the document;
Sources	use terms that might be included in a standard subject index; use important terms from title and their synonyms and related terms; consult citations from relevant literature; consult published glossaries and thesauruses.

Instructions for Manuscript Documentation Unit

Table 3

Category 4.12:	Compilers and Generators	
-- language	procedure oriented	
	context free	language
	formal	language
	graphic	language
	interactive	language
	programming	language
	syntax specified	language
	meta	language
-- compil	compiler	
	compilation	
	compiler-compiler	
	metacompiler	
	incremental compilation	
-- translat	translation	
	programming language translator	
	general translator	
	meta language translator	
	translator	
-- pars	parser	
	parsing	
-- system	compiler writing	system
	translator writing	system
-- processor	programming language	processor
	macro instruction	processor
	general	processor
	meta	processor
	meta language	processor
	macro	processor
syntac (x) --	syntax specified language	
	syntax	
	syntax directed	
	syntactic analysis	

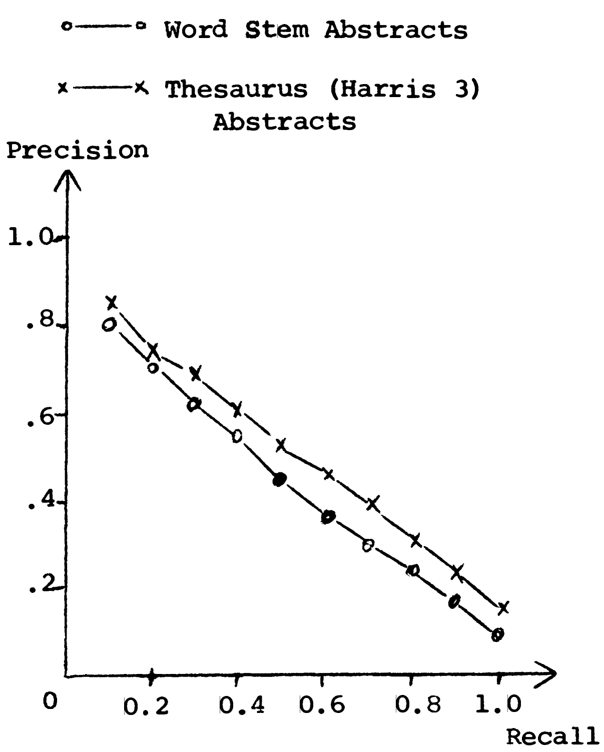
Common Vocabulary for Category 4.12
(6 documents)

Table 4

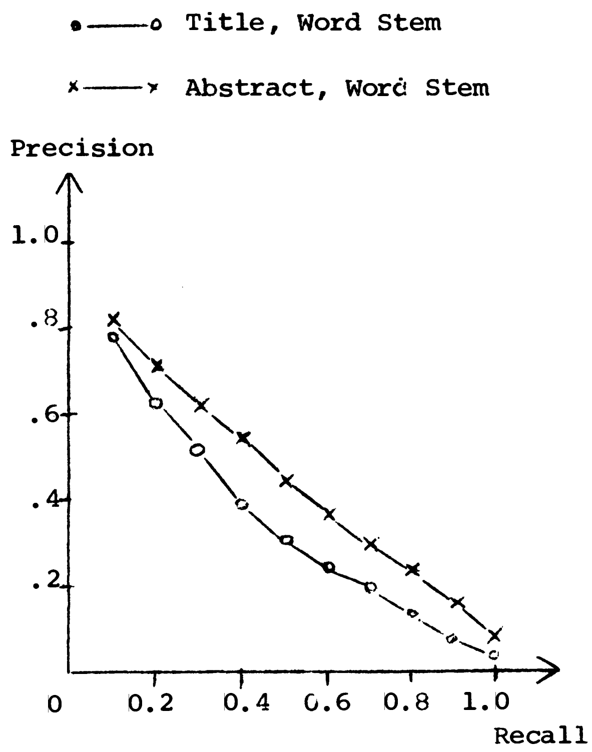
Common Categories	Document Numbers	Key Phrases
4.41 Utility Programs, Input-Output 6.35 Input-Output Equipment	Document 24: (4.41, 6.35)	blind communication blind programming aid Braille Braille output Braille teletype Braille communication tactile terminal tactile computer communication
	Document 29: (3.81, 4.41, 6.35)	telephone communication transmission telephone errors error correction
1.52 University Courses	Document 17: (1.1, 1.51, 1.52, 1.59)	computer appreciation courses for liberal arts students survey courses beginning programming course content drop-out rates in computer courses college versus precollege teaching and social responsibility
	Document 23: (1.52, 2.45)	computing centers research instruction utilization expenditures support higher education

Illustration of Lack of Common Vocabulary
(broad categories; inappropriate indexing)

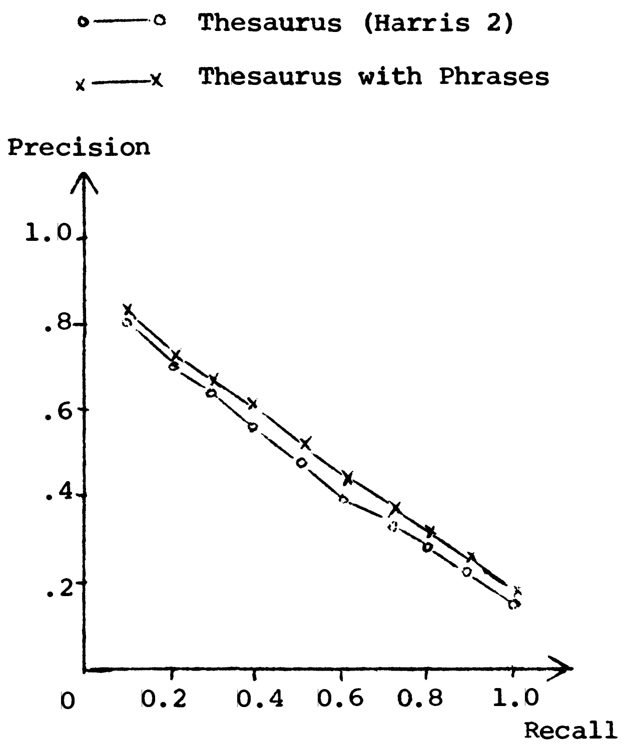
Table 5



a) Stem-Thesaurus



b) Title-Abstract



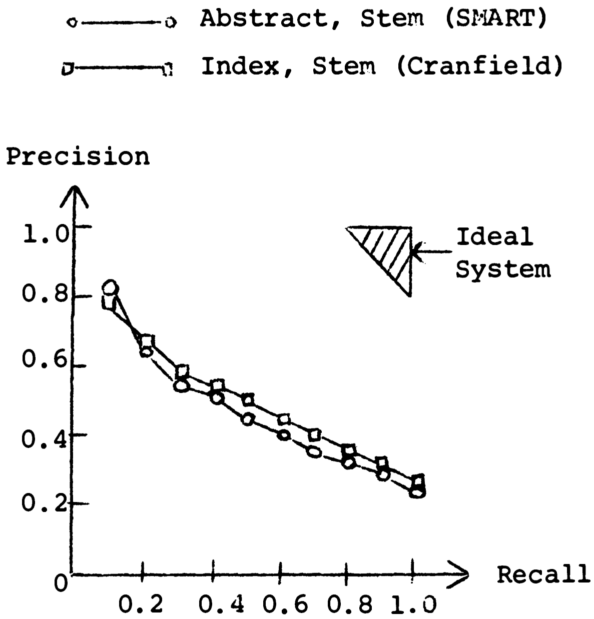
c) Thesaurus-Phrases

Evaluation Measure Precision at Recall Pt.			Statistical Significance (Sign Test)
R	Stem	Thes	
0.1	0.816	0.895	0.454
0.3	0.632	0.718	0.075
0.5	0.459	0.568	0.007
0.7	0.308	0.459	0.001
0.9	0.169	0.302	0.000

Evaluation Measure Precision at Recall Pt.			Statistical Significance (Sign Test)
R	Title	Abstract	
0.1	0.800	0.816	0.302
0.3	0.527	0.632	0.012
0.5	0.317	0.459	0.011
0.7	0.207	0.308	0.020
0.9	0.085	0.169	0.001

Evaluation Measure Precision at Recall Pt.			Statistical Significance (Sign Test)
R	Thes	Phrases	
0.1	0.826	0.837	0.267
0.3	0.640	0.661	0.286
0.5	0.452	0.481	0.839
0.7	0.338	0.356	1.000
0.9	0.219	0.236	0.248

Comparison of Text Processing Systems (IRE-3 Collection)
(780 documents-34 queries)
Fig. 1



Evaluation Measure Precision at Recall Point			Statistical Significance (Sign Test)
R	Abstract	Index	
0.1	0.824	0.804	0.701
0.2	0.652	0.658	1.000
0.3	0.558	0.591	0.728
0.4	0.509	0.550	0.860
0.5	0.452	0.517	1.000
0.6	0.414	0.451	0.736
0.7	0.380	0.403	0.868
0.8	0.343	0.365	0.743
0.9	0.300	0.323	0.256
1.0	0.255	0.280	0.065

a) Recall-Precision Graph

b) Significance Output

Comparison of Automatic Word Stem Process with
Manual Indexing

(Cranfield Collection - 200 abstracts, 42 queries)

Fig. 2

ACM Author Instructions for Manuscript Documentation Unit

(Effective January 1, 1968)

The usefulness of articles published in ACM periodicals is greatly enhanced when each paper is accompanied by information which insures proper indexing, classification, retrieval, and dissemination. To this effect there will be published with each article appropriate documentation information consisting of

a) descriptive title;

b) author names and addresses;

c) informative abstract;

d) content indicators of two types:

i) appropriate key words and key phrases,

ii) category numbers from *Computing Reviews* (CR);

e) citations to the relevant literature.

The above information must also be furnished by authors at the time of submission of the original manuscript as a separate *Manuscript Documentation Unit*, and its completeness and accuracy are taken into account by referees and editors in reviewing the manuscript (i.e., the manuscript for publication should be submitted in the format as previously customary, except with the addition of Key Words and Phrases and CR Categories following the Abstract; the manuscript for publication and the manuscript documentation unit are two distinct entities, transmitted at the same time).

The following suggestions may be useful in preparing descriptive titles, content indicators, and informative abstracts.

Descriptive Title. Use a specific and informative title to tell accurately and clearly what the document is about. Choose title terms as highly specific as content and emphasis of the paper permit. Typically, a title might contain six to twelve words. Avoid special symbols and formulas in titles unless essential to indicate content. "Cute" or "clever" titles are unhelpful and should not be used.

Informative Abstract. The abstract should consist of short, direct, and complete sentences. A reading of the abstract should serve in some cases as a substitute for reading the paper itself. For this reason, the abstract should be *informative*. Typically, its length might be between 150 and 200 words. The abstract should state the object of work, summarize the results, and give the principal conclusions and recommendations. It should state clearly whether the focus is on theoretical developments or on practical questions, and whether subject matter or method are emphasized; also whether the article is a review, or survey, or tutorial. Original data in the form of new data or methods of procedure should be included and emphasized according to their importance. The title need not be repeated. Work planned but not done should *not* be described in the abstract.

Content Indicators. Two types of content indicators are to be assigned: category numbers from the classification schedule used by *Computing Reviews*, and free choice key words and key phrases consisting of English language words.

To assign the *CR category numbers* consult the latest *CR* classification schedule, herewith or published in *CR*, May-June 1967 issue, pages 302-303. Use as many category numbers as may be applicable. If possible, specify your interpretation of the "miscellaneous" or "general" categories if these are used. The following category numbers might, for example, be applicable to a manuscript dealing with sorting techniques: 3.74 (searching), 4.49 (miscellaneous utility programs), 5.31 (sorting).

In listing *key words* and *key phrases* to be used for indexing, put yourself in the place of the person who is looking for information in your paper, and write down all the words that he might use in searching an index. If you have a technical thesaurus available, such as the *IFIP-ICC Vocabulary of Information Processing* [North-Holland Publishing Co., Amsterdam], consult it. Also, for helpful suggestions for alternate key words consult the citations to the relevant literature. The key words and key phrases used should be as precise as possible and hopefully unambiguous in their particular context. Typically ten to fifteen words or phrases might be used. The following additional guidelines may be of help:

a) use important terms from the title; include also their synonyms, related words, and words of higher or lower generic rank;

b) use English nouns, or noun-noun and noun-adjective combinations; do *not* use prepositions; do *not* use sequences of more than three words; do *not* use hyphens except if the hyphenated parts are *always* treated as a single unit;

c) use specific terms whose meaning is generally accepted in the computer field; do *not* use broad catchall terms (such as "computer," "automatic," "machine," "system," "discussion," "description"); do not use private terms or acronyms that may not be generally known;

d) do *not* use negative terms stressing what your paper does not do; emphasize the positive content and contribution.

Indexing Example.

Minsky, M. L., Artificial Intelligence. *Sci. Amer. 215*, 3 (Sept. 1966), 247-260.

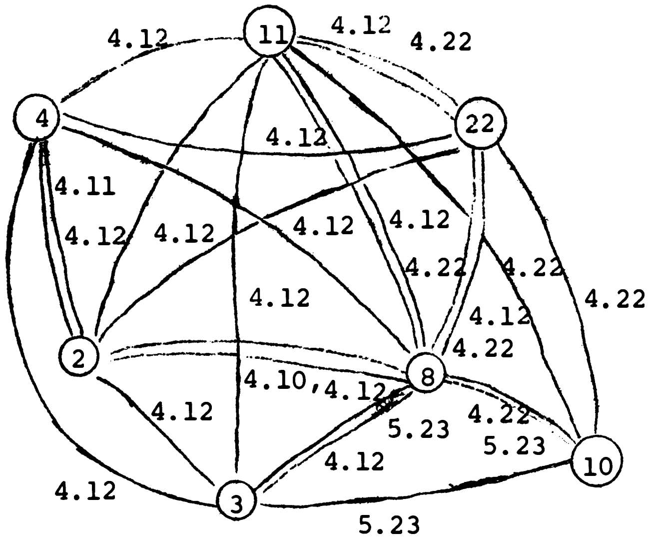
Abstract: This is a tutorial paper dealing with some of the achievements and problems in artificial intelligence. A number of well-known applications, including checker playing, picture matching, and problem solving in high school algebra, are used to demonstrate that machines can be made to exhibit intelligence. Typically, the programs include methods for setting up goals, making plans, considering hypotheses and recognizing analogies.

Key Words and Phrases: artificial intelligence, intelligence h[illegible]ing, heuristic procedures, heuristic programming, game playing, checkers, feature recognition, pattern recognition, pattern matching, list processing, picture processing, question answering, language processing, language analysis, problem solving, hypothesis testing.

CR Categories: 3.60, 3.61, 3.62, 3.63.

Manuscript Documentation Unit

Fig. 3



a) Document Graph

Documents	2.11	2.3	3.74	4.1	4.10	4.11	4.12	4.13	4.20	4.22	4.9	5.23	5.24
	Category Numbers												
2					✓	✓	✓		✓				
3							✓					✓	
4			✓			✓	✓				✓		
8				✓	✓		✓			✓		✓	
10										✓		✓	✓
11	✓	✓					✓	✓		✓			
22							✓			✓			
	Computing Milian		Applications	Programming								Mathematics	

b) Assignment of Category Numbers

Sample Document Network

Fig. 4