CLASSIFICATION, RELEVANCE, AND INFORMATION RETRIEVAL

D.M. Jackson

TR 70-80
November 1970

Department of Computer Science
Upson Hall
Cornell University
Ithaca, New York 14853

CLASSIFICATION, RELEVANCE, AND INFORMATION RETRIEVAL

D. M. Jackson

Technical Report

No. 70-80

November 1970

Department of Computer Science
Upson Hall
Cornell University
Ithaca, New York 14850

Classification, Relevance, and
Information Retrieval

D. M. Jackson[1]

---

1. To appear in Advances in Computers, 11 (1970)

1. Introduction

To treat the relationship between classification, relevance, and information retrieval, it is first desirable to consider briefly some of the more important events that have taken place in the field of information retrieval. A brief account of the field follows.

One of the earliest references to the problem, which has since become central to information retrieval, was made by Bush [8] in 1945. Commenting on the inadequacies of contemporary library classifications and of the resulting tendency to neglect existing information on research work, he remarked: "Even the modern great library is not generally consulted: it is nibbled away at by a few ... our ineptitude at getting at the record is largely caused by the artificiality of systems of indexing." Since that time, work has progressed in a number of different directions and has achieved a number of goals, all falling within the general area of information retrieval. All have the common problem of extracting or retrieving information, although for different purposes, from a corpus of natural language text or from a prepared set of data. Darlington [13], for example, shows how certain English problems can be translated into logical terms and treated subsequently by formal methods. Bobrow [5] deals with the linguistic problems that arise while analyzing those English statements which describe mathematical relationships. The "word problems" are translated directly into algebraic expressions. Simmons [60] finds answers in an indexed natural language text by means of grammatical dependencies. Raphael's [51] SIR (Semantic Information Retrieval) demonstrates how conversational and deductive abilities may be obtained by using a model which represents semantic content from a variety of subject areas. Green's [22] BASEBALL answers questions posed in ordinary English about the data in store and determines what information is given about the data being requested. An important distinction must be made here between these systems and those in which the computer acts primarily as a repository for records of information (medical or insurance records, for example) and delivers the correct record once a suitable tag (the patient's or customer's name) has been provided. In the systems under consideration, the linguistic relationships between the request and the set of data being interrogated are examined to establish a response. In the others, no such relationships are determined -- the tag addresses directly the information to be retrieved. Throughout this chapter, the term "information retrieval" will be taken as synonymous with "document retrieval" (<u>sensu</u> Cooper [11]) -- that is, a system which refers the user to documents which may be of interest to him, by examining a request which indicates his area of interest. This is, therefore, contrasted with retrieval systems that deliver specific items of interest in the subject area of interest in response to a user's request. This usage agrees with that of Salton [57] and Cleverdon [10] and many others.

Since the importance of the problem was recognized, mainly in the United States, a number of retrieval systems have been designed, suggested, and even used. The use of high speed digital computers profoundly affected the field and permitted the mechanization of certain stages of the retrieval process. Computers were most effective in the stages which were purely routine in character and which could be performed equally well without the participation of the "intelligent" mind. The recording of the removal or return of books, the issuing of "overdue" notices, the calculation of fines and the verification of the identity of a library user and his status to access restricted regions of the library are examples of such stages. The University of Pittsburgh, for example, has mechanized some of these stages and the system is now in constant use.

The development of large capacity storage media accessible to computers has also resulted in innovations in the area of information retrieval and this has had its attendant problems. The difficulties of scanning large interdependent files of information at a cost comparable with the present manual systems, the updating and verification of the integrity of the information, and the preserving of the complete system for recovery after a computer breakdown are all problems to which people have addressed themselves.

Attention has not, however, been confined to these areas alone. The systems referred to earlier aim at identifying information contained within written text in order to perform some appropriate action. Few systems deal with English text of unrestricted structure. The counterpart in information retrieval is the isolation of information contained within documents[2] written in natural language to specify what the text is "about," and the subsequent use of this information to assist in the retrieval of only relevant[3] documents in response to the user's request. In doing so, the computer has been used to perform a task which has been the subject of long investigation by librarians. Maron [37] speculates on the direction in which this type of work may progress:

> To correctly classify an object or event is a mark of intelligence; a mark of even greater intelligence is to be able to modify and create new and more fruitful categories, i.e., to form new concepts. (Perhaps one of the really dominant characteristics of the intelligent machine will be that of creating new categories into which to sort its "experiences".)

The information retrieval systems that have been suggested and used involve the following stages, some of which may be wholly or partially mechanized:

(a) The documents added to the system are examined to determine information about them which will facilitate their retrieval.

---

2. Document here is used in the generic sense, namely, a a piece of natural language text.

3. Relevant is used, for the moment, informally. The dictionary definition "bearing on, pertinent to, the matter in hand" will suffice in the present context.

(b) The requests are examined to determine the information essential for the retrieval of documents relevant to them.

(c) The information provided by (a) and (b) is then compared or matched in such a way that the document collection may be partitioned into two sets, namely, the set of documents relevant to the request and the set of documents not relevant to the request. Alternatively, the documents are ordered according to their relevance to the request, according to some scale of relevance.

(d) Access to the relevant documents is provided by divulging their reference numbers.

In connection with these, two principal subjects of research arise, both of which have been studied extensively in the last two decades. These are:

(1) What information should be extracted from documents and requests so that algorithms may be designed to achieve (c) above?

(2) Which stages may be mechanized, and how?

More recently, a third area of research has arisen. That is the following.

(3) In what ways may retrieval systems be evaluated in terms of their ability to retrieve relevant documents in response to users' requests?

There have been two main approaches to the first stage of information retrieval, namely the processing of documents to be added to the system. These approaches are described in the following two sections. Question (3) is discussed in Section 1.4.

1.1 The Classification Approach

Documents are classified or grouped into a particular order, according to a predetermined classification scheme. The documents may be grouped physically on shelves. This is the principle upon which most conventional library classifications are based (e.g., Dewey, UDC). This principle of classification makes a number of presuppositions. The following are taken from Shera [59] and adapted for discussion of the classification of knowledge in general.

(a) Linearity of subject arrangement. This is necessary in order physically to arrange books on shelves.

(b) Complete covering of all knowledge.

(c) Meaningfulness of terms and meaningful differences between terms.

(d) Provision for a specific place for every topic or field.

Such classifications encounter difficulties which do not seem to be soluble, even by revision of the classification which is used. Thus, when it is unclear into which of several categories a document should be placed, rules have to be devised to permit the selection of one class in favor of another. These are essentially arbitrary for they are exterior to the general principle of the classification scheme. Without duplicate copies of documents or an adequate set of cross-references, documents which relate to the subject area of a particular class may, nevertheless, not appear in that class and therefore may be missed in retrieval. Thus the limitations of the traditional classifications are that they are inconsistent in their organization and therefore excessively complex in structure, and are unable to adapt to the advance of knowledge.

The way in which such a classification is established owes a great deal to the Aristotelean method. Such an approach is _a priori_ in character. A method is given for the analysis of classes by enumerating their contents and distinguishing each from the others. Thus, starting with the _genus_, a series of _differences_ is stated. The _species_ are then listed according to their different characteristics or _properties_. A close parallel may be drawn between the classifications used in libraries and the development of botanical classifications. In botanical classification, the major subdivisions were determined by considerations which, on an _a priori_ basis, seemed to be of paramount importance in contemporary science. There was no attempt at experiment, and as a result, there was considerable diversity of opinion about the criteria which were important in making the major divisions. A number of different classification schemes arose. Caesalpino (1519-1603) paid chief attention to the fruits of the plant. Tournefort

(1656-1708), however, based his classifications on the character of the corolla. Linnaeus (1707-1778), on the other hand, regarded the variations of the sexual organs of plants as the main consideration in making the divisions. The systems grew in complexity, as have the library classifications of the present day, as exceptions to them were found and recorded in the classification. As Adanson (1727-1806) observed, there appeared to be little consistency between the classifications.

The difficulty of fulfilling requirements (a) - (d) in Section 1.1 is plain. Even in the case of (c), the problem is not straightforward, for it is not always clear what constitutes a meaningful difference. The following, taken from Diogenes Laertius [15], is sufficient to illustrate the point briefly:

> Plato had defined Man as an animal, biped and featherless, and was applauded. Diogenes plucked a fowl and brought it to the lecture-room with the words "Here is Plato's man." In consequence of which there was added to the definition, "Having broad nails."

A more serious criticism of the _a priori_ method of library classification arises in connection with requirements (b) and (d), namely complete _covering_ of all knowledge and the provision for a _specific_ _place_ for every topic or field. It is as if all the objects of the universe had been distributed into lists or parcels, and a common name given to the objects of each list until all the general names in language had been invented, thereby covering all knowledge. To decide whether a certain general name may be truly predicated of a particular object, it is necessary, as it were, only to read the role of objects which have the name to see whether the object appears among them. Mill [40] remarks: "So absurd a doctrine will be owned by nobody when thus nakedly stated;..." He adds: "General names are not marks put upon definite objects; classes are not made by drawing a line round a number of given assignable individuals. The objects which compose any given class are perpetually fluctuating." Thus if the meaning of a general name is the set of things which have the name, no general name has a fixed meaning. In discussing the meaning of general names he asserts:

> The only mode in which any general name has a definite meaning, is by being the name of an indefinite variety of things; namely, of all things, known and unknown, past, present, or future, which possess certain definite attributes. When, by studying not the meaning of words, but the phenomena of nature, we discover that these attributes are possessed by some objects not previously known to possess them (as when chemists found that the diamond was combustible) we include this new object in the class. We place the individual in the class because the proposition is true; the proposition is not true because the object is placed in the class.

In information retrieval one is confronted with the problem of assigning documents to a set of categories which are not necessarily exclusive or independent. Maron _et al_ [37] clearly point out the difficulty:

> The problem arises because the categories are not defined extensionally. That is, a category is not defined by enumerating its members (documents), but, rather, the situation is reversed. Based on some more or less clear notion of a category, we must decide whether or not to assign a document to it.

Needham [42] remarks in this connection:

> If we are to assign a document to a class automatically, we must have (a) a list of facts about the classes which will make ascription possible; (b) an algorithm, usually some sort of matching algorithm, to tell us which class best suits a document. Given a classification like the UDC, it is not at all obvious that (a) and (b) exist, or even, if they can be found. (a) and (b) imply a degree of uniformity about the classification which may just not be there.

1.2 The Indexing Approach

The purpose of indexing is to describe or characterize the content of a document to ensure its retrieval by the appropriate request. Consideration is confined exclusively to indexing methods in which relational operators between terms are not used. A detailed comparison of the other class of methods is given by Soergel [62]. One of the most commonly used methods was "multiple aspect" indexing. Here a document was described by a list of terms. The terms were regarded as unordered and unrelated. A request, processed in much the same way, was compared with a document to determine whether it should be retrieved or not in response to the request. A decision to retrieve or not was taken, possibly after a number of further operations, by determining whether the term list of the document contained the terms of the request. As such, the systems were realizable on punched cards and, in principle, terms common to the request and the document were detected by sensing holes in identical positions on two cards by superimposing the cards. There were two main approaches. The "Uniterm" endeavored to restrict the thought required of the indexer to a minimum. The indexer recorded the terms which he considered were characteristic of the document, and in doing so a term vocabulary was gradually accumulated _a posteriori_ from his lists. Originally, no effort was made to record synonyms or near-synonyms. The vocabulary grew in size without any firm internal structure. The virtue of such a system was in its simplicity, and its merit lay in its attempt to reduce work on the part of the indexer to a minimum. For general use, however, the system proved to be ineffectual.

The "thesaurus" approach attempted to remedy the ills of the Uniterm system by controlling the internal structure of the vocabulary. The method is generally associated in its original form with Luhn [35]. A term vocabulary was determined in advance, and terms were included in it for their effectiveness in describing the subject matter of the document collection as a whole. The indexer was instructed to select terms from this vocabulary that were appropriate to the documents he examined. This approach proved to be considerably more successful than the Uniterm system. Its disadvantages, however, lay in the difficulty of extending the system to deal with the accumulation of documents which were gradually added to the collection,thereby altering its original character. Thus new terms needed to be added to the term vocabulary. Some of the old terms needed to be more specific in their meaning to distinguish between the new concepts introduced by the new documents. To achieve this it was necessary to replace the offending terms by a number of new terms. To include these changes, the complete document collection had to be re-examined to reassign to them the terms of the new vocabulary. The resemblance between a structured vocabulary of this type and the present day technical thesaurus is strong. At the time, however, no such thesaurus existed. Roget's _Thesaurus_, with its complete lack of technical terminology, was inappropriate for document collections concerned with specific research topics in technical fields.

A similar attempt to exercise some control over the term vocabulary was provided by the _association map_ or _library lattice_ of Doyle [16]. It was successful in dealing with synonyms and provided a means of replacing arbitrary terms by a more specific term or a more general term. Giuliano [20] remarked about their use:

> Automatic index term techniques are needed to improve the recall of relevant information, to enable indexers and requestors to use language in a more natural manner, and to enable retrieval of relevant messages which are described by different index terms than those used in the inquiry.

The lattice is a partially ordered set and the ordering relation was that of the generic to the specific. Thus synonyms were identified in the lattice with the same nodes. Near-synonyms were represented by nodes that were close together in the sense that their least upper bound was a few arcs above them. The replacement of a term or a set of terms by a more general term related to it in the lattice had the anticipated effect of increasing the number of documents retrieved but ran the risk of retrieving documents which were not relevant. The general term comprised the word senses which were isolated and distinguished in the dependant nodes lower in the lattice. The replacing of a term by a more specific term had the reverse effect. Fewer documents tended to be retrieved, but the relevance of these was hopefully enhanced. The price of this was that a number of documents relevant to the request were lost. This inverse relationship characterizes all attempts at automatic retrieval.

Needham [43] has proposed and experimented with an ordering relation which differs from the "generic-specific" relationship mentioned above. Needham's relationship was expressed informally by the slogan: "If, when you ask for a document about A you cannot reasonably complain if you get a document about B, then A should be above B in the ordering." The ordering relation was therefore more clearly related to the retrieval process in that the relation itself involved a rudimentary assessment of the relevance of documents to subject categories. The ordering relation, nevertheless, still depended strongly upon subjective decisions. What, for example, constitutes a "reasonable complaint"? In the use of this ordering relation may be seen the beginning of a "mechanistic" approach to the construction of term classifications for information retrieval. In such an approach, attention is focused primarily on reproducing the expected responses by the system as requests are presented to it, without the anterior assumptions of the utility, in retrieval, of particular linguistic relationships between terms. The determination of the linguistic character of the classes of a classification, developed as the result of a purely mechanistic approach, would, therefore, be posterior to the construction of the classification.

A number of methods have been proposed to generate automatically the lattice dependencies between the items of the term vocabulary. Some of the techniques have involved a syntactic analysis of the text with a subsequent manual screening to select the unique correct analysis for each sentence. The relationships between nouns located in this analysis were then computed with reference to the co-occurence of terms within sentences and within syntactic structures. Such a method is only partially automatic since it involves the manual stage of selecting the correct syntactic structure. In addition, the syntactic analysis of large quantities of text with which methods must be capable of dealing in the future, may scarcely be contemplated unless the methods are greatly refined.

The measurement of the co-occurence of pairs of terms within documents has been used by Stiles [67], and by Borko and Bernick [7] in factor analytic methods. These methods have aimed either at the assignment of documents to subject categories in the case of Borko and Bernick, or at the reformulation of users' search requests by Stiles. The reformulation of the request expanded it by the addition of statistically associated terms to refine the original request. Factor analytic methods have proved effective in many of the reported experiments. Their principal drawback is that they are not amenable to the examination of large document collections. Their computational requirements as determined by the existing eigenvector methods make such procedures impracticable for such a purpose.

The statistical methods that have been outlined briefly above represent a significant departure from the indexing methods of the 1950's. The earlier methods were concerned more closely with the behavior and meaning of words within a more extensive area than that covered by the particular document collection being used in retrieval. They were concerned with the meaning and use of words in natural language as a whole. The whole experience of the indexer, as a person well versed in the language in which the document was written, was brought to bear when he indexed the collection. The statistical techniques, at least those concerned with term association, deal only with the vocabulary of the document collection and the ways in which the words relate within the document collection.

Again, a parallel may be drawn between the developments of the last two decades in library classification and the gradual recognition, in post-Linnaean botanical classification, of the importance of observation. The dictatorship of the Aristotelean method, with its notion of arriving at the truth by reason unaccompanied by experiment, was on the decline during Bacon's lifetime. In its stead followed the extensive and detailed examination of plants and the painstaking recording of botanical observations. Classification was to be achieved by the comparison of as many features of the plants as possible. Adanson [1] wrote of this method: "La vrai fisique des Plantes est donc celle qui considère les raports de toutes leurs parties & qualités, sans excepter une seule; elle réunit toutes les Plantes en Familles naturels & invariables,..." Sneath [61] has discussed Adanson's contribution to the development of modern taxonomy in detail.

The use of word frequency statistics for the analysis of the content of documents is ascribed to Luhn [35]. His work, begun in 1957, was concerned with the analysis of single documents, and it was not until two years later that the statistical approach was extended beyond the single document stage. The use of computers and the popularity of coordinate indexing, in which terms were assigned to documents without differentiating between the relative importance of the terms in describing the content of the documents, led to "associative indexing." Maron and Kuhns [38] recognized the inherent unsoundness of undifferentiated terms in coordinate indexing and explained:

> The correspondence between the information content of a document and the set of tags is not exact as it is difficult to express precisely the subject content of a document by means of one or more index terms. There is no strict one-to-one correspondence between the set of index terms on the one hand and the class of subjects they denote on the other hand. We say that there is semantic noise in the index terms.

The semantic noise occurs, they claim, because the meaning of the terms is dependent upon their context and function within the documents. "The grammatical type, position and frequency of other words help to clarify and specify the meanings of a given term." The use of Boolean combinations of the index terms of requests has been suggested for reducing semantic noise. If, for example, two terms are connected conjunctively, the subject area is narrowed and more closely specified. The disjunction of terms, on the other hand, broadens the scope of the request. Thus, by the appropriate use of logical combinations, the retrieval of irrelevant documents or the missing of relevant documents would, hopefully, be avoided. The conjunction of terms in the request, although it decreases the chance of obtaining irrelevant documents, has the effect of increasing the chance of missing relevant documents. The converse holds for the disjunction of terms in the request. Verhoeff [75] has demonstrated this behavior, and the resulting inapplicability of Boolean functions in general to information retrieval systems.

The associative methods rest on the assumption that sufficient information may be obtained from the documents by counting word and word-pair frequencies to represent adequately the information content of the documents. The assumption, if it is appropriate, relieves us of the detailed and costly work of examining the context and function of terms within documents in order to resolve the "semantic noise" referred to by Maron and Kuhns. Maron _et al_ [39] defend the principle as follows:

> Whereas the semantical relationships are based solely on the meaning of terms and hence are independent of the "facts" described by these words, the statistical relationships between terms are based solely on the relative frequency with which they appear and hence _are_ based on the nature of the facts described by the documents.

> Thus, although there is nothing about the meaning of the term "logic" which implies "switching theory," the nature of the facts (viz. that truth functional logic is widely used for the analysis and synthesis of switching circuits) "causes" a statistical relationship. Another example might be the terms "information theory" and "Shannon"...

Doyle [18] remarks in the same connection: "When pairs of words co-occur strongly in text it should not be surprising to find the members of the pairs associated in the minds of the authors of books in which they occur." Thus, in short, the basis for associative indexing is the assumption that strongly co-occuring terms are likely to be conceptually related.

The associative methods led to "probabilistic indexing", in which a weight is given to each term assigned to a document giving the probable relevance of the index terms to the document. Dissimilarity measures between documents, significance measures for index terms, and closeness measures between the terms were required, together with the _a priori_ probability distribution data. Classes of documents were retrieved and ranked according to their "computed relevance numbers". Thus, in addition to avoiding the disadvantages of coordinate indexing with its use of undifferentiated terms, Maron tried to make the retrieval process a natural extension of his statistical approach. However, there appears to be no clear distinction between indexing and classification. Thus Maron [37] states: "This enquiry examines a technique for automatically classifying (indexing) documents according to their subject content." Later he states: "This approach to the problem of automatic indexing is a statistical one. It is based on the rather straightforward notion that individual words in a document function as clues, on the basis of which a prediction can be made about the subject category to which the document most probably belongs." The method relies on the existence of established subject categories to which the documents are to be assigned. The categories are not generated automatically. This subject classification is an _a priori_ one, with the result that the method is limited by the refinement and sophistication of the classification.

1.3 The Evaluation of Retrieval Systems

Although the work done during the last two decades on automatic subject classification and automatic indexing has been considerable,[4] little progress has been made in the evaluation of retrieval systems until recently. The problems associated with evaluation, however, have been recognized for some time. The earliest comparative experiment, using a coordinated uniterm index, was reported by Gull [23] in 1956. The need for more adequate evaluative procedures has been made apparent by the progress in automatic classification and indexing techniques. Research in information retrieval

4. See, for example, the state-of-the-art report by Stevens [66].

will soon reach the point where, hopefully, sufficient experimental evidence will have been collected about existing classification and indexing methods to provide the basis for the establishing of a formal theory. Indeed, a theoretical approach may be regarded as a natural stage in the evolution of retrieval systems, starting from the early days of Bush's MEMEX and the mechanical approach, and progressing through the experimental stage of the late fifties and early sixties. The aim of such a theory would be to distinguish between "good" and "bad" retrieval practice in the design of retrieval systems, and to allow some sort of estimate to be made of the operating characteristics of a retrieval system once its design had been specified. Such a theory is not yet available, although the research associated with the SMART system and with the Cranfield Project is progressing in this direction. Evaluation serves two important purposes. First, it provides a means of choosing between a number of possible systems on the basis of the excellence of their performance as retrieval systems. Second, it provides a means of judging the effect on performance of variations in the techniques used in retrieval or the variations in experimental design. This is essential if research in information retrieval is to progress.

Cleverdon [10] has suggested a number of criteria of importance in the evaluation of retrieval systems.

(a) The ability of the system to present all relevant documents (i.e., recall);

(b) The ability of the system to withhold all non-relevant documents (i.e., precision);

(c) The interval between the demand being made and the answer being given (i.e., time);

(d) The physical form of the output (i.e.,presentation);

(e) The effort, intellectual or physical, demanded of the user (i.e., effort).

Cleverdon has further argued that of these (c) and (d) offer the least problem since both may be improved by more sophisticated or more appropriate hardware. The evaluation of (e) may be achieved by the straightforward observation of a number of users of the system. The main intellectual difficulties of evaluation reside in providing suitable methods of estimating (a) and (b), namely, recall and precision.

In a retrieval system which is to be operable on an industrial scale, an evaluation measure would supposedly strike an economic balance between the quality of the performance, judged in terms of recall and precision, and the cost of achieving this. Supposedly, some reduction in quality might be tolerable provided that the cost of maintaining the system decreased appreciably. Such consideration, however, may be deferred until the task of designing a system that exhibits good performance in terms of recall and precision has been accomplished. Accordingly, throughout this chapter, the performance of a retrieval system is discussed solely in terms of recall and precision. Their appropriateness for evaluating retrieval performance is strongly defended by Cleverdon: "The unarguable fact, however, is that they [i.e., recall and precision] are fundamental requirements of the users, and it is quite unrealistic to try to measure how effectively a system or subsystem is operating without bringing in recall and precision."

An evaluation measure gives an estimate of the system's ability to retrieve all documents relevant to the users' requests. Ideally, for effective comparisons to be made between systems, this estimate must be characteristic of the system itself, and therefore independent of the characteristics of the user population the system serves. There are a number of difficulties associated with this. The evaluation of a retrieval system requires that a set of requests be available, together with a set of decisions giving the relevance of each document in the collection to each request. The response of the system to the set of requests is compared with the expected response, and the agreement or disagreement between these is used to evaluate the performance of the system. Provided that the comparison is made over enough requests, it may be supposed that evaluation is independent of the particular choice of test requests. Performance of a given system may change with the subject area covered by the document collection, as the SMART system has shown. The evaluation also depends on the individual decisions concerning the relevance of each document of the collection to each request, and these decisions are essentially subjective. Mooers [41] is emphatic on this point: "There is no absolute 'relevance' of a document. It depends upon the person and his background, the work and the date. What is not relevant today may be relevant tomorrow." In the last analysis, relevance may only be defined in terms of what the user considers appropriate as responses to his request. Relevance, nevertheless, for all its disagreeable and apparently arbitrary properties remains central to any discussion of evaluation. As Doyle [19] aptly and succinctly remarked: "Relevance is a thought crutch; with it we may think inaccurately about the retrieval problem, but without it (or something better) we couldn't think at all."

A number of attempts has been made to examine the question of relevance and its applicability and suitability as a basis for the determination of retrieval system performance. Cuadra and Katter [12] remark: "Relevance judgements have been used as a basis for measures designed to evaluate the effectiveness of information retrieval systems.... There is reason to believe, however, that as ordinarily obtained, they [i.e., the relevance judgments] may be unreliable and sensitive to a number of conditions of measurement that have not been carefully controlled in previous evaluation studies." In their study, a number of criteria or "implicit use orientations" are defined, on the basis of which subjects (judges) are invited to assess the relevance of documents to requests. Not surprisingly they report: "It was found that the particular use orientations assumed by the judge has a marked effect on relevance judgments." The use orientations enabled the judges to adopt a particular "assigned point of view" about relevance. If a judge is instructed, for example, to determine relevance according to the "stimulative value"[5] of documents then his judgments would, in all likelihood, differ from those where he judges according to the "hard-factual"[6] nature of documents retrieved. A request offered to a retrieval system seldom specifies in sufficient detail the actual intention or motivation of the user in submitting the request. However, the difficulties raised by Cuadra and Katter might be lessened, although not eliminated, if the "use orientation" were incorporated explicitly into the request. To assess relevance of documents to request according to first one "user orientation" and then according to a second, for example, is to submit two, possibly substantially different, requests. Cuadra and Katter conclude:

> These findings support the belief that relevance scores are very likely to be artifacts of particular experimental instructions and conditions and should not be offered or viewed as absolute numbers. This conclusion does not invalidate experiments in which a given group of judges makes relevance appraisals for a specific system. However, it does call in question any comparitive evaluations (between systems or sub-systems) in which the attitude of the judges about

---

5. Stimulative value, i.e., wants articles that will have "idea-stimulative" value for himself.

6. Hard-factual, i.e., fact-oriented and very chary of "theory and speculation"; is interested in "hard results."

the intended use of the materials was not considered and controlled, either experimentally, or statistically.

Resnick and Savage [52], as do Cuadra and Katter, support the use of relevance as the basis of evaluation for a given system. In one of their experiments they have set out "...to compare the relative inter- and intrasubject consistency of humans when judging the relevance of documents to their general interests on the basis of (a) citations, (b) abstracts, (c) index terms and (d) total texts." Briefly a number of randomly chosen subjects, informed in the topic areas, were directed to choose from a number of documents, randomly selected from a set of internal publications, those documents which were relevant to their interests. Their decisions were based on the complete document, citations, abstracts, or index terms. The experiment was repeated with the same judges and materials after the lapse of a month. The results of the two experiments were compared to determine

(1) The ability of humans to judge consistently the relevance of items to their general interests.

(2) The effect on these judgments of different kinds of lexical material (documents, citations, abstracts, and index terms).

(3) The ability of humans to recall previous judgments of relevance.

The results showed that "...under the conditions of this experiment, humans are consistent in their judgment of relevance and that this consistency is independent of the kind of material on which this judgment is based, except in the case of abstracts. This anomaly associated with abstract judgment appears to have no simple explanation." They conclude, "The high consistency of the human judgments observed in this study may put to rest those criticisms of the reliability of evaluations of text processing systems based on such judgments." They also agree with Cuadra and Katter about the artificiality of relevance numbers giving the degree of relevance of a document to a request, and observe: "What does remain as a problem is a procedure for the quantification of the results of human judgments.... The scores [i.e., relevance scores], however, still cannot be shown to lie on a metric scale, and we are not yet justified in assuming that a score of 50 is twice as high as a score of 25. The problem awaits further analysis."

Taube [72] also agreed that, although judges may be selfconsistent in their judgments of relevance, it is not clear that they are consistent among themselves. He remarks:

> When relevance is used to characterize a document or characterize the relationship between a document and the subjective requirements of a searcher, its meaning is as clear as any other psychological predicate; that is, we may accept the supposition that the searcher uses the term relevance with fair consistency in referring to his own responses. When it comes to assessing the consistency of usage for all searchers, we are on much shakier ground.

Taube, like Cuadra, Katter, Resnick, Savage and others has warned against current practices in the measurement of relevance. He has objected that, while relevance is widely agreed to be a subjective notion, reasonably consistently assessed by a single judge but admitting considerable variation between judges, it has nevertheless been calculated in many experiments to a degree of arithmetic precision which is unreasonable and inappropriate. In addition, since relevance assessments may not be consistent between users, relevance-recall values may not be taken as characteristic of the retrieval system itself. In this connection Taube asserts,

> When...discussion shifts to the <u>relevance-recall</u> ratio, relevance is assumed to characterize or describe, not psychological states, or the relation of a particular document to a particular question, but the system as a whole. It is supposedly an indexing or IR system as a total complex which is characterized or ranked according to its relevance and recall performance. It may be said that the system's performance is measured by subjective responses....

Thus, if the evaluation of performance is sensitive to changes in the assessment of relevance of the type which may occur between competent and informed judges, then certainly the evaluation is not characteristic of the system. Taube has also questioned the use of recall and precision on procedural grounds. "A subjective response may be trusted to determine that a certain percentage of items retrieved are not relevant, but ... cannot establish a percentage of recall because the subject does not know what other documents there are in the collection which might be relevant to his question. This criticism cannot be made if sufficient care has been taken before evaluation to ensure that th relevance of each document to each request has been examined.

Recent studies by Salton [57] have clarified some of Taube's criticisms. The purpose of the study was to determine the effect of variations in relevance assessments on the average recall and precision values used in measuring retrieval performance. Briefly, four sets of decisions were obtained from a number of judges giving the relevance of each document to each of the requests. Two of the sets were the set of relevance decisions obtained by the originators of the requests, and the set of relevance decisions obtained by a person other than the originator of the requests. The decisions were dichotomous, and indicated only the relevance or nonrelevance of a document to a request. In addition, the crierion on which the judges were instructed to assess relevance was strict. A document was to be judged relevant to a request only: "if it is directly stated in the abstract as printed, or can be directly deduced from the printed abstracts that the document contains information on the topic asked for in the query." Three different retrieval procedures were used and these were ranked according to their retrieval effectiveness computed for each of the sets of relevance judgments in turn. Salton asserts, "...it appears reasonable to conclude not only that the performance methods are ranked in the same order, no matter which of the four sets of relevance judgments is used, but also that the actual performance differences resulting from differences between author and nonauthor judgments are negligible." The relative insensitivity of the evaluation of the retrieval methods to relatively large differences in the relevance judgments is ascribed to the calculation of recall and precision averages over the set of requests. Salton condludes:

> ...it can be stated that, if the relevance assessments obtained from the query authors used in the present study are typical of what can be expected from general user populations of retrieval systems, then the resulting average recall-precision figures appear to be stable indicators of system performance which do in fact reflect actual retrieval effectiveness.

Similar conclusions have been reached by Cleverdon [9] in connection with the ranking of a number of indexing languages according to their retrieval effectiveness. Cleverdon concludes, "Insufficient analysis has been done to make any detailed comments, but it is clear that the relevance decisions did not significantly effect the comparative results of Cranfield II."

2. TERM CLASSIFICATION IN INFORMATION RETRIEVAL

2.1 Intersubstitutability, Co-occurrence, and Stability

The merits of indexing by "uniterms" has already been discussed in Section 1. Both the simplicity with which uniterms may be used and the lack of sophistication required of the indexer in assigning them to documents lessen the cost of the initial stage of document processing. Of the methods reviewed in Section 1 for representing the information content of the document collection, the use of uniterms is at least economically the most practicable on a large scale. The complete absence of any structure between them, however, makes them on their own entirely unsuitable for information retrieval. The "thesaurus" approach discussed in Section 1, however, provides a structure between terms which may profitably be used in retrieval. The difficulty of extending the thesaurus or lattice as the document collection increases in size and the necessity of reindexing the complete collection if terms are added to the vocabulary, are the principal obstacles to the thesaurus in information retrieval. Provided that these disadvantages may be overcome, it is appropriate to examine a retrieval system which combines the simplicity of the uniterm approach with the organized structure of the thesaurus. Accordingly the function of the library lattice will be examined in greater detail to determine how the difficulties may be resolved.

Suppose that for a particular document collection a lattice has been constructed. The system may be realized on punched cards, a card being set aside for each term in the vocabulary. Each card also contains an indication of the terms above and the terms below the term represented by the card. The system is operated in the following manner. For each term in a request presented to the system the appropriate cards are selected and superimposed. The documents which are to be retrieved are indicated by the presence of a hole in each card in the same position. This may be sensed by optical or mechanical means. The output of such a system may be increased by replacing some or all of the cards representing request terms by the cards which correspond to those terms immediately above the request terms in the lattice. These terms are available on the cards corresponding to the request terms. "Above" and "below" are interpreted in terms of the ordering relation which induced the lattice. In a lattice in which the ordering relation is from the "specific" to the "generic", the replacement of a term by a term above it in the lattice corresponds to a broadening of the request. The replacement of a term by a term below it in the lattice corresponds, on the other hand, to a narrowing of the request. All documents indexed under a particular term will also be indexed under the less specific term above it in the lattice.

Suppose for example that Fig. 1 is a fragment of a lattice giving the relationship between the terms t_1, t_2, t_3, and t_4 and t_5. A request which contains the term t_1

may be broadened by replacing it by t_4. Not only are the documents containing t_1 retrieved, but also those documents containing the term t_2, which is also directly below t_4 in the lattice. Thus, considering the replacement by t_4, t_1, and t_2 may be regarded as <u>intersubstitutable</u> since the output of the system will be unchanged by replacing t_1 by t_2 in the request.

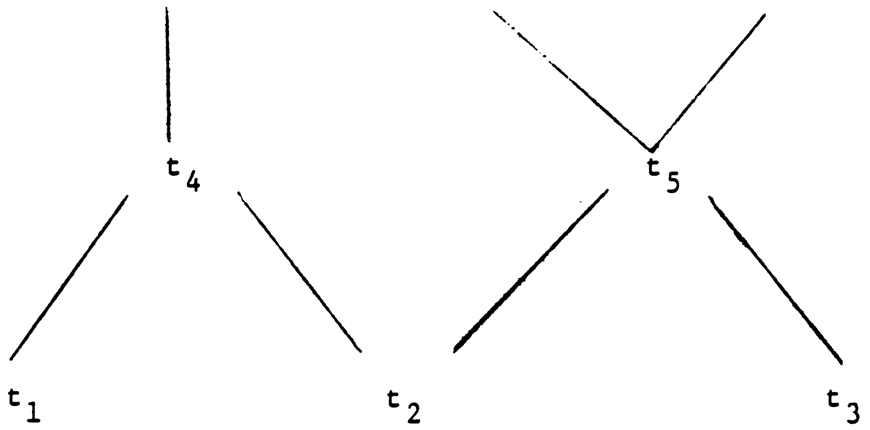

Fig. 1. Fragment of a lattice

The retrieving of a document is unaltered if one of its terms t_1 is replaced by another of its terms t_2. Thus, considering the retrieval of a specific document, the terms which occur together in the document are intersubstitutable. A specific pair of terms are intersubstitutable over the whole document collection provided only that the terms occur together in documents of the collection. This is unlikely to occur in practice and this requirement of intersubstitutability must therefore be relaxed. Instead, complete intersubstitutability over the whole document collection may be approximated by those terms which occur together in a significant number of documents of the collection. More generally, groups, rather than pairs, of terms are required which comprise terms showing a marked tendency to occur together in documents. Pairs of terms taken from the same group will be assumed to be intersubstitutable, although this assumption may only be supported in practice if the tendency to co-occur is sufficiently strong for each pair of the group. The construction of the lattice has therefore been reduced to a classification problem which, hopefully, is more tractable.

The groups of terms whose members show a marked tendency to co-occur resemble, in usage, the groups of terms that are closely related in the lattice, although there may be little resemblance between the constitutions of the groups. The groups of terms may be regarded as descriptors and the complete set of descriptors as a classification of the term vocabulary. The descriptors are, as it were, one level removed from the document text itself. Words extracted from the document text itself are more closely related to the text than a descriptor, for a descriptor may never appear in the text. The similarity in function between groups of terms sharing the same upper bound in the lattice and groups of terms whose members show a strong tendency to co-occur may be seen from Fig. 1. Just as a document or a request containing t_1 may be expanded to increase recall by substituting for t_1 the nearest term above it given by the partial ordering relation, so t_1 may be replaced by the descriptor corresponding to the class containing t_1. In addition, the other terms in the same class as t_1 are involved in the substitution in the same way that t_2 was involved by replacing t_1 by t_4. The document and request are, in effect, expanded by the addition of all terms which occur in the same class as terms in the document or terms in the request. Thus the descriptors, defined by the classification of the term vocabulary, are used for coordinate indexing, although the search techniques which will be used in retrieving documents will be different from those used, for example, with the UDC.

From the original term descriptions of the documents the relative frequency of co-occurrence of terms within documents may be derived. On the basis of this information about the interrelation of terms it is proposed to construct a classification. A document, may not, however, have a unique index description in that the decision to assign a term or keyword to a document, particularly with the uniterm system, rests with the discretion of the indexer. Although two indexers, indexing the same documents, may produce descriptions which agree with each other substantially, it is nevertheless likely that there will be points of disagreement. Accordingly, it is necessary that two term classifications derived from the same document collection, indexed separately by two different indexers, should not differ substantially. That is, the classification technique must be stable with regard to small changes in the document descriptions. Thus, for example, it would be unreasonable to suppose that the addition of a single document to the collection would result in a serious reorganization of the classification. Borko [6] makes a covert reference to the stability requirements of term classifications when he remarks:

> One of the claims for mathematically derived classification systems is that the categories so derived are descriptions of the documents used in the analysis. However, if the categories prove to be so unique that they describe only the one document set and no other, they would be of little value.

Needham and Sparck Jones [45] make a similar observation: "We want to be able to update our classification: we should be able to add new terms without radically altering the system if there are only a few of them." This requirement is a constraint on the classification technique and providing the technique is stable with respect to small changes in term description, significant differences in term classifications, should they occur, are attributable only to the inadequacy of the indexing. Assuming, however, that the documents have been indexed adequately, the use of term classifications in retrieval has the important property of reducing the effects of bad or faulty indexing decisions.

Classification is a process in which the individual differences between the objects to be classified are suppressed in order to derive more general properties which are characteristic of groups of objects. Two questions therefore arise in the present context:

(1) Has too little information been lost, in that the classification permits distinctions to be made between terms which uselessly distinguish between certain documents?

(2) Has too much information been lost, in that the classification does not enable appropriate distinctions to be made between certain documents?

If too much information has been lost then the performance of the retrieval system will doubtless suffer, and this will be shown in the evaluation of the system. If, on the other hand, too little information has been lost, then effort has been wasted in making distinctions which are redundant for the purposes of retrieval. It is important that this waste of effort should be avoided in a system that is intended to be viable on a large scale. Questions (1) and (2) may be summarized by adapting the terminology of the requirements of Section 1; that is, are the classes obtained <u>meaningful</u> and are there <u>meaningful</u> <u>differences</u> between classes which enable the appropriate, and only the appropriate, distinctions to be made between terms?

These questions may be answered in full only after the classification has been set up and documents have been retrieved from the collection in response to requests submitted to the system. Evaluation of the response of the system based on the ideas of Section 1 may be used to determine the adequacy of the classification. A partial answer may be given to the first question, however, on the basis of the type of information that will be used in the construction of the term classification. A classification constructed from the co-occurrence of terms in documents is less likely to make redundant distinctions between terms than a human classifier since there would be no information on the basis of which these could be made. The human classifier, on the other hand, has available additional sources of information about the terms and, on the basis of this, he is able to make distinctions between terms which in fact serve no useful function in indexing the collection.

2.2 Retrieval Characteristics of Term Classifications

The use of the relative frequency of occurrence of terms within documents has been arrived at by considering only the intersubstitutability of terms in the library lattice. Maron et al [39] and Doyle [18], among others, have supported the use of relative frequency of co-occurrence for determining groups of semantically associated words. Terms that occur together in the same document may not have any obvious semantic relationship, and, accordingly, the linguistic character of the classes derived from term co-occurrence will not be obvious. The terms within a class will not be more strongly related than "topically" related, in that they may meaningfully be used in connection with the same topic area within the subject area of the collection. Accordingly, it is reasonable to suppose that groups of intersubstitutable terms will consist of those terms that relate to topic areas within the field covered by the document collection. The classification of the term vocabulary into groups of terms whose members show a strong tendency to occur together in documents enables document and request descriptions to be expanded to include terms related to the topic area of the request. For each document or request there are, therefore, two specifications: the first using manually assigned terms from the term vocabulary, and the second using automatically assigned descriptors defined by the constructed term classification. The extent to which a document relates to a request is determined by the degree to which the document and request specifications conform with each other. Provided that the assignment of terms to documents of the collection is adequate for characterizing their content, the greater the conformity between document and request specifications then the greater the relevance of the document to the request. The degree to which a document and request conform may be determined by a function called the <u>match function</u>. A number of different match functions may be defined, each of which represents a different way of uniting term and descriptor specifications to improve the performance of the retrieval system. The match functions must all exhibit, however, the same general behavior with respect to the term and descriptor specifications of documents and requests as was described above, namely, the greater the agreement between term specifications and between descriptor specifications then the greater the value of the match function. Given a particular document and a particular request, there are four specifications -- two term specifications and two descriptor specifications -- that may be used by the system to determine whether or not the document should be retrieved in response to the request. The probability that a descriptor will be shared both by the request and the document is higher than the probability that there will be a shared term, since a descriptor is shared provided at least one term from the request belongs to the same class as at least one term from the document. Accordingly, as has been asserted above, descriptors have the effect of increasing the recall of the system. The same assertion may be made on intuitive grounds, for the expansion of the term descriptions by using descriptors enables the system to retrieve those documents which relate to the same topic area as the request. Thus the use of a term classification in information retrieval may be regarded as a recall device, to compensate for the failure in term matching between document and request.

A term classification may also serve as a device for improving the precision of the retrieval system. Suppose that the classification is such that classes are permitted to overlap freely. A term may, therefore, belong to more than one class. The different senses of a term or the different contextual uses of a term may therefore be identified by determining the term's different class affiliations. The class membership of a term specifies the term's affiliations to the topic areas of the document collection. The senses are indistinguishable if simple term descriptions of documents are used. The same conclusion has been reached by Vaswani [73] who states: "For the purposes of document indexing a word might usefully occur in more than one cluster to allow it to be clustered with sets of words with which it may associate in different meanings.... The optimum degree of overlap may have to be found by experiment."

Both recall and precision are affected by the sizes of the classes of the classification. If the classes are large, then the probability of a descriptor occurring both in the document specification and in the request specification is high, since a descriptor will be shared if there is at least one term in the request which belongs to the same class as at least one term in the document. Precision is low for large classes since the document and request have, in effect, been augmented by a large number of terms. For smaller classes, the probability of a shared descriptor decreases so that the recall decreases accordingly. Precision, on the other hand, may be higher since the document and request specifications have not been augmented by terms which may result in the retrieving of nonrelevant documents. As has been seen above, a term classification may be used both as a recall and as a precision device. The problem of constructing classifications for information retrieval is to ensure that they are increased simultaneously.

2.3 Limit on Attainable Retrieval Performance

The approach to information retrieval discussed above enables improvement to be made in the performance of a retrieval system by utilizing a classification of the term vocabulary. The classification is constructed automatically from information which is assumed to be adequate in representing enough of the content of the documents for the purposes of information retrieval. The retrieval performance is to be assessed on the basis of the evaluation methods outlined in Section 1.3 in which the responses of the system to a set of requests are compared with the expected responses of the system.

The retrieval system, however, may be such that, given the general specification of permissible term classifications and document-request match functions, the required response of the system to the base set of requests may never be realized. That is, no choice of classification and no formulation of the document-request match function, within the general design of the retrieval system, may achieve perfect performance. The highest performance attainable by the system lies somewhere below this ideal. Part of the difficulty lies in designing a system with sufficient flexibility to accommodate all the vagaries in relevance judgments made by users of the system. Salton [57] has made the observation that "The gap between the complete human search and a perfect search...appears to be due to ambiguities in query formulations and to the difficulties of reconciling the user's view of a subject area with the subject classification provided in a given document collection. This latter gap may never be bridged by any search and retrieval system likely to come into existence in the foreseeable future." This is not, however, the only difficulty. There may be a limit on the attainable performance imposed by the design of the retrieval system irrespective of the set of relevance judgments, and this may be a significant contribution to the observed gap between actual performance and perfect performance. There has been little reference in the literature to the existence of a limiting value of achievable performance for a retrieval system, and, certainly, no methods have been proposed for determining it. It is important, at least in information retrieval research, to know the limits imposed on performance by the actual design of the system. Detailed discussion of this and associated problems is deferred to Section 4.

The properties of term classifications required in information retrieval may now be summarized.

(1) The classification is to be based on information relating to the relative frequency of co-occurrence of terms within documents.

(2) The classes comprise those terms which show a marked tendency to occur together in the same documents.

(3) The classes may overlap freely and are arranged nonhierarchically.

(4) The membership of terms to classes is not quantified by probabilistic weights. A term either belongs to a class or does not belong to a class.

(5) The classification must be stable with respect to small changes in the term descriptions.

Methods for generating classifications conforming to these requirements are now considered.

3. THE CONSTRUCTION OF TERM CLASSIFICATIONS

3.1 General Requirements

The type of classifications discussed are unusual in that classes are intentionally allowed to overlap. The role of overlap between classes of terms as a possible precision device in information retrieval was referred to in Section 2. In many applications, however, it is a requirement that the objects to be classified should be assigned to disjoint classes, and the clustering and hierarchical methods for achieving this comprise the main bulk of classification techniques. In plant taxonomy, for example, where the objectives are somewhat different since categorical differences between the objects to be classified are of primary importance, Williams and Dale [76] are skeptical about the use of overlapping classifications. Techniques appropriate to these applications are given by Sokal and Sneath [63]. Nevertheless, the techniques to be described below for constructing overlapping classifications have been applied, with some success, by Parker-Rhodes [46] and Parker-Rhodes and Jackson [47] to mycological data containing ecological and meteorological descriptors.

In information retrieval, classes of terms must be formed in such a way that they may be used to achieve an increase in retrieval performance. There is, however, no _a priori_ reason for claiming that a particular classification in which the members of the classes are related in some specific way, is better than any other classification of the terms. The value of a classification is determined empirically by evaluating the effect on retrieval performance when the classification is incorporated into the system. Thus, in constructing classifications for information retrieval, the Platonic concept of error in the assignment of terms to classes is entirely absent. That is, there is no prior classification which automatic techniques are intended to reconstruct in detail; and there is, therefore, no basis for the probabilistic assignment of terms to classes to give the extent that properties of particular terms in the vocabulary are inferable from a knowledge of the classes to which each belongs. Sutcliffe [68, 69] has attempted to replace the Platonic concept of error by comparing, statistically, a large number of classifications of the same population to determine the typical and atypical allocation of objects to classes. The stringent conditions under which observations on the objects must be made so that atypical class allocation may be adopted as an indication of error, make this approach inapplicable here.

Certain classifications suggest themselves as devices for improving retrieval performance, and it is quite consistent to assert that these improve retrieval performance provided that such an assertion is phrased in terms of a hypothesis that is susceptible to testing. It may be supposed, for example, that a classification of the term vocabulary into groups of synonyms can be used to improve the performance of a retrieval system beyond that obtained with other classifications. Gotlieb and Kumar [21] propose another classification based on the *a priori* (pre-assigned) semantic relationships between terms rather than on the co-occurence of terms within documents. In their scheme, the "closeness" of two terms is determined from the sets of all immediate semantic relatives of each terms (i.e., all terms in the vocabulary which reference each term and which are referenced by each term). Their method requires a cross-reference mapping giving the "see also" references, as is provided, for example, in the Library of Congress' *Subject Headings* [34].

The assertion that a particular classification improves performance for a particular collection may be tested directly by evaluating the effect of the classification on retrieval performance, and the validity of the assertion for document collections in general may be determined provided sufficiently many document collections are examined. However, the success with which terms have been arranged into groups of synonyms must be determined by inspection of the classes and intuitive assessment of the constitution of each class. As a relation between terms, synonymy is quite strong and, indeed, methods have been proposed for isolating groups of synonyms automatically by Sparck Jones [64] and by Lewis et al. [33]. Topical relatedness, suggested by the arguments of Section 2, is however less definite. The answer to the question whether two terms are topically related -- that is, the two terms may relate to the same topic area -- as opposed to synonymously related, is less obvious. In probabilistic classification, the probability of the assignment of terms to a class gives the reliability of the information conveyed by asserting that the term belongs to the class. Topical relatedness is not sufficiently definite a relationship in the first place to permit the calculation of probabilities with any degree of accuracy. Accordingly, a classification in which the assignment of terms to classes is nonprobabilistic is more appropriate for dealing with topical relatedness.

The classifications which we consider consist of classes of terms that show a marked tendency to occur together in the documents of the collection. The classes formed by this criterion will, hopefully, specify topic areas within the field covered by the collection, and accordingly, terms that appear together in the same class will be said to be topically related. The criterion on which classes are formed is a weak one in that classes are to be composed of terms that <u>tend</u> to co-occur; co-occurence between every pair of terms in a class is not demanded. The criterion is therefore less restrictive than for cliques in which the classificatory criterion must be exhibited between every pair of objects belonging to a clique. A clique is a maximal complete subgraph of a graph in which the presence of an arc between a pair of nodes indicates that the nodes are related. The degree of relatedness may be specified by the strength of the arc. A further requirement of the classification is that the classes be self-generated in the following sense. Classes are defined as sets of terms which satisfy the classificatory criterion, and the complete vocabulary must be explored to find such classes. It is inadmissible to specify a set of possible topic classes initially and subsequently to distribute the terms to the most appropriate of these since, first, these classes must be treated as unknown, second, significant yet intuitively obscure classes would not be located, and third, the number of classes required is unknown. Such procedures, for example, Sebestyen's [58], might, however, be adopted as an economical means of updating an existing classification to avoid complete reclassification of the term vocabulary as the document collection expands. In terms of the general statement of the "classification problem," we want to construct classes of objects such that objects more strongly resemble objects of the same class than they do objects not in the same class. The objects are terms and the degree of similarity is determined by the relative frequency of co-occurrence of terms within documents. Accordingly, a matrix giving the similarity between pairs of terms of the vocabulary is required. This commits the order of the classification procedure to be at least n^2 in the number of objects to be classified. Methods exist which avoid the use of such a matrix. Doyle's [17] algorithm, for example, which locates clusters of documents, does so by comparing each document with a profile of the class to determine whether or not to assign the document to it. It is, however, necessary to specify in advance the number of clusters required in the final classification. Doyle's algorithm has been shown not to converge in certain cases. Dattola [14] has improved Doyle's algorithm and has obtained convergence. The convergence of a general class of classification procedures, akin to Dattola's has been demonstrated by Needham [44]. Dattola's algorithm is of order n log n in the number of objects to be classified. The conditions imposed on the determination of similarity between terms will now be given.

The objects to be classified are described or defined in terms of measurements or observations made upon them and the information is characteristically conveyed by specifying the qualitative[7] or quantitative[8] attributes which the objects possess. A great deal of discussion in taxonomic literature has been devoted to the problems arising from the interdependence of attributes. Suppose, for example, that the binary attributes "black" and "white" were used in describing the objects of a population. Then, provided that these attributes were relevant to the population, the value of one of the attributes would determine the value of the other.

---

7. These are sometimes called "qualities." The attributes may be two-state, indicating the presence or absense of the quality; or may be multistate. The attribute "leaf-shape," for example, in which linear, obovoid and pandurate must be distinguished, is three-state. Note the independence of the states.

8. For example, "leaf-length," possibly averaged over a suitable sample.

A less obvious example is given by the use of "carnivorous" and "ungual." Here the dependence is not a logical one. Sokal and Sneath [63] are of the opinion that dependencies between attributes which are known without recourse to analysis must be eliminated before the objects are classified. Williams and Dale [76], while admitting that such dependencies may dominate the analysis, consider that their use in classification is a matter of taste -- the sole criterion is the interest or otherwise of the user. Thus, since the dependence between "carnivorous" and "ungual" has been established, subsequent users can afford to be more sophisticated in their choice of initial attributes. The problem corresponds in information retrieval to the presence of terms which co-ccur frequently compared with the frequency of occurrence of either term. The acceptability of the dependencies must be determined by their effect on retrieval performance. There is evidence that they have a deleterious effect. The dependence, however is not a logical one.

3.2 Constraints on Classification Procedure

Suppose that x and y are two objects of a population U and suppose that $d(x,y)$ is the similarity between x and y. Then the following conditions are imposed:

$$d(x, y) = d(y,x) \quad \text{for all } x, y \text{ in } U \tag{1}$$
(symmetry)

Blank page - xeroxing error

$$d(x,y) \geq 0 \qquad \text{for all } x,y \text{ in } U \tag{2}$$

$$d(x,y) = 0 \qquad \text{if } x = y \text{ for } x,y \text{ in } U \tag{3}$$

(indistinguishability of identical objects)

$$d(x,y) = 0 \qquad \text{only if } x = y \text{ for } x,y \text{ in } U \tag{4}$$

(distinguishability of nonidentical objects)

$$d(x,z) \leq d(x,y) + d(y,z) \qquad \text{for all } x,y,z \text{ in } U \tag{5}$$

(triangle inequality)

A system which satisfies (1) - (5) is called Euclidean. A system in which (5) is not everywhere true is called a quasimetric. A system in which (4) is not everywhere true is called a pseudometric (or a semimetric). Systems have been examined in which (5) is replaced by the more restrictive condition:

$$\text{(5')} \quad d(x,z) \leq \max(d(x,y),d(y,z)) \qquad \text{for all } x,y,z \text{ in } U$$

Systems in which (5') is imposed are called ultrametric. Taxonomic hierarchies using ultrametrics have been discussed by Johnson [29], and by Jardine et al. [28]. A number of functions satisfying some or all of (1) - (5) have been proposed for information retrieval, for example, by Kuhns [30], and for wider applications, for example by Ball [2].

A number of "distance coefficients" which satisfy some or all of (1) - (5) have been extensively applied in ecology and anthropology. Typical among these coefficients are those of Pearson [49], Mahalanobis [36], and Rao [50]. These coefficients are intended for use when the descriptions of the objects to be classified themselves result from the statistical determination of the mean and variance of a character or attribute over a sample of all instances of the object in the population. The counterpart in information retrieval is the determination of the "distance" or similarity between index terms, taking account of the variability in the assignment of terms to documents by a sufficiently large number of indexers independently indexing the collection. The uniterm indexing scheme was considered because of its simplicity and cheapness to apply in practice. The possibility of repeated indexing of the same collection would add grossly to the cost of obtaining the initial index descriptions, and accordingly cannot be considered.

Observations on the terms of the vocabulary consist of the names (call numbers) of those documents to which the terms have been assigned. Since the documents of the collection are indexed using uniterms, each document may be characterized by a vector whose components, one or zero, indicate whether or not a particular term has been used in indexing the document. The complete collection may therefore be represented by a rectangular array D, the _i_th row of which gives the index description of the _i_th document. $D_{ij} = 0$ indicates that document i has not been indexed by term j, while $D_{ij} = 1$ indicates that it has. Since it is the terms that are to be classified by their co-occurrence in documents, the terms are the objects of the universe (the document collection) while the documents are two-state or binary attributes. D is called the document array or the document matrix. Co-occurrence of terms within documents may be determined from the matrix D quite simply. Consider two terms t_j and t_k. Then the frequency of co-occurrence of the terms is given by

$$|D_{\circ j} \wedge D_{\circ k}|$$

where $D_{\circ j}$ is the jth column of the document array D. This will be called the mutual association of t_j with t_k. The self-association of t_j is defined as the number of documents in which it occurs and is given by

$$|D_{\cdot j}|$$

These quantities arise naturally in discussions of computational methods for generating similarity matrices. The relative frequency of co-occurrence of terms within documents may now be defined as

$$s_1(t_j, t_k) = |D_{\cdot j} \wedge D_{\circ k}| \;/\; |D_{\cdot j} \vee D_{\circ k}| \tag{6}$$

This is a quasimetric and varies between 0, when t_j and t_k never co-occur, and 1, when t_j and t_k always co-occur. The function, which arises naturally in information retrieval, is ascribed to Tanimoto [71]. It is biassed in favor of positive occurrence of terms in documents. The common non-occurrence of terms in a document does not contribute toward the similarity of the two terms. As has been already seen in Section 2, intersubstitutability of terms is based on actual occurrence of terms in documents. A symmetrical form of the function may, however, be defined. A function with similar properties is the correlation coefficient, defined by:

$$s_2(t_j,\ t_k\ ) = \left[\frac{|D_{\cdot j}\ \wedge\ D_{\cdot k}|^2}{|D_{\cdot j}|\ \cdot\ |D_{\cdot k}|}\right]^{1/2} \qquad (7)$$

It is clear that other functions may be defined for calculating the similarity between pairs of terms based on their frequency of co-occurrence within documents. Account may be taken, for example, of the variation in the numbers of terms assigned to documents. The functions defined above, however, have the substantial practical advantage of making the manual inspection of classes, member by member, a relatively straightforward matter. As an example of these functions, suppose that in a collection of 5 documents, t_1 appears in documents 1, 2, and 4 while t_2 appears in documents 1, 2, 3, and 5. Then t_1 and t_2 are given vectorially by

$$t_1 = (1,1,0,1,0)$$

and

$$t_2 = (1,1,1,0,1)$$

Therefore $s_1(t_1,\ t_2) = 0.4$ and $s_2(t_1,\ t_2) = 1/3$.

The classification procedure which is needed must construct classes of intersubstitutable pairs of terms. Intersubstitutability has already been expressed in terms of the co-occurrence of terms within documents, from which a value of the similarity between the two terms may be computed. These values, for all pairs of terms in the vocabulary, form a matrix which provides the information about the terms from which classes are to be constructed. The following are constraints generally placed on classification procedures:

(a) The classification must be well defined. That is, application of the algorithm must supply a single result.

(b) The classification must be stable. That is, the classification must not be grossly affected by small changes in the data.

(c) The procedure must be unaffected by a permutation of the names of the terms. That is, the procedure must be independent of the labeling of the terms.

(d) The procedure must be independent of scale. That is, the procedure must be unaffected by multiplication of the similarity matrix by a positive constant.

The need for stability has been remarked already, for it enables the classification to be reasonably insensitive to faulty indexing of the document collection. Independence of scale is an obvious requirement since the scale of the similarity function is essentially arbitrary. The stronger condition of invariance under a monotonic transformation of the similarity function has also been suggested.

For example, a number of methods have been proposed in which distances or similarities between objects are largely irrelevant. The methods rely on the inequalities between pairs of objects. Similarity functions induce a partial ordering on the set of objects to be classified. A classification is constructed such that a partial ordering derived from it agrees as closely as possible with the partial ordering induced on the data. The measurement of aggreement is given by Benzécri [3]. The methods are accordingly insensitive to monotonic transformations of the similarity functions with the result that, according to Benzécri [4], "...the host of indexes enumerated by S. and S.[9] is essentially reduced to two, which differ mainly (and do definately differ) according to the part played by NN matches..." An NN match is a match in which a given attribute is recorded as absent in both objects compared. Such a match may or may not be allowed to contribute toward the total similarity between the two objects. Benzécri distinguishes the two cases. An overview of these methods is given by de la Véga [74].

Independence of labeling requires that a classification should not be dependent on the order in which the terms are treated, and is demonstrable only in a classification in which all classes which satisfy the criterion are enumerated. For the algorithm to be proposed, it cannot be shown that all of

---

9. Sokal and Sneath.

the classes are found. A permutation of the names of the terms would result in a classification which, although sharing a substantial number of classes, would differ in a number of classes. The failure of (a) and (c) are related. Relabeling enables more classes to be found, and examination of all permutations corresponds to inspecting all subsets of the terms to determine which satisfy the classificatory criterion. It is clearly impracticable to examine all subsets since there are too many of them. Provided a large enough number of classes is found the failure of (a) and (c) is unimportant.

It is not easy to give a precise meaning to "large enough." Certainly, the least number of classes required to distinguish all of the terms is the binary logarithm of the number of terms in the vocabulary. A rough guide is provided by requiring that the classes, used as descriptors, should provide a more economical description of the objects than the original attributes.

3.3 The Classificatory Criterion

It is now possible to give a more detailed specification of the classification criterion; that the classification should consist of classes such that each term resembles terms of the same class more than it does terms not in the same class. Two things are needed. First, a means of measuring the overall similarity of a term to a class; second, a means of deciding to which, among a number of classes, a given term has the greatest similarity. Care must be taken in specifying how this decision is made to avoid presupposing the existence of a fixed number of classes among which the term is to be distributed. This presupposition violates the requirement that classes be self-generating. The difficulty may be resolved quite simply as follows. Suppose that A is a set of terms and that B is the complement of A with respect to U, the term vocabulary. Suppose further that A is a putative class of the classification. The construction of A progresses by deciding, for each term, whether to admit a term into the putative class or to reject it; that is, by deciding whether to assign the term to A or to B. Thus, membership to B is by default and corresponds to rejection from A. In general, B will not be a class of the classification.

Suppose that A is a set of terms. Suppose also that x is a term and that s(x,A) is the similarity between x and the class A. It is not clear that any strict rules may be laid down for the design of s(x,A). Lance and Williams [32], in their treatment of hierarchical systems, have examined the effect of object-set and set-set similarity functions on local geometry. They argue that the spaces defined by such functions do not necessarily have the same properties as the space defined by the interobject distances. Certain classification strategies will behave as though the space in the immediate vicinity of a group has been distorted (that is, contracted or dilated). <u>Space contraction</u>

corresponds to an apparent movement of a group on formation toward some or all of the remaining objects. Space dilation corresponds to a movement away from other objects. Lance and Williams warn against the use of space-contracting systems. These concepts also apply to clustering systems which optimize intragroup homogeneity given by Lance and Williams [31].

However, a number of points are clear. The similarity between x and A is the same as the similarity between A and x. The similarity between a term and a set must reduce, in the case where the set consists of one term, to the similarity between the two terms. This is equivalent to the definition of compatible by Lance and Williams [32] in their work on hierarchical systems. A compatible strategy is one in which the similarity between two sets reduces, in the case where the sets are single membered, to the similarity between the two members. Also, the similarity between x and A must increase when another term with non-zero similarity to x is added to A. Finally, the similarity between x and A must depend only on the individual similarities between the terms of A and x. This requirement arises naturally in Section 2. Lance and Williams [31, 32] refer to strategies with this constraint as combinatorial. A non-combinatorial strategy, on the other hand, is one in which the attributes of the objects must be reconsulted in determining similarity. The distinction is primarily concerned with the computational aspects of classification and is of little theoretical interest except where, for example, the comparison of certain attributes requires human judgment.

An appropriate function which fulfills these requirements is

$$s(x,A) = \sum_{y \in A} s(x,y)$$

Analogous requirements may be placed on the similarity between two sets A' and A, and an appropriate definition is:

$$s(A', A) = \sum_{x \in A'} s(x, A)$$

where A' and A are disjoint and

$$s(A,A) = \frac{1}{2} \sum_{x \in A} s(x,A) \qquad \text{[each } s(x,y) \text{ counted only once]}$$

Suppose that A and B are sets of terms defined as above. Suppose also that q is the scale of the term-term similarity function so that $0 \leq s(x,y) \leq q$. The scale of similarity will be normalized by choosing $q = 1$. The following relationships hold for any such A,B:

$$s(A,A) + s(A,B) + s(B,B) = \text{const.} \tag{8}$$

for the term vocabulary and the document collection. This may be shown as follows:

$$s(A,A) + s(A,B) + s(B,B) = \frac{1}{2} \sum_{x \in A} \sum_{y \in A} s(x,y) + \sum_{x \in A} \sum_{y \in B} s(x,y)$$

$$+ \frac{1}{2} \sum_{x \in B} \sum_{y \in B} s(x,y)$$

$$= \frac{1}{2} \sum_{x \in A} \left(\sum_{y \in A} + \sum_{y \in B} \right) s(x,y)$$

$$+ \frac{1}{2} \sum_{y \in B} \left(\sum_{x \in A} + \sum_{x \in B} \right) s(x,y)$$

by symmetry for term-set similarity.

$$= \frac{1}{2} \sum_{x \in A} s(x,U) + \frac{1}{2} \sum_{y \in B} s(U,y)$$

$= s(U,U)$ by symmetry for set-set similarity

= constant for the term vocabulary and the document collection

= S, say.

Let $s^*(A,A)$ be the value of $s(A,A)$ when all terms in A are maximally similar to each other. Then

$$s^*(A,A) = \frac{1}{2} N_A (N_A - 1) \qquad (9)$$

where $N_A = |A|$ since

$$s^*(A,A) = \frac{1}{2} \sum_{x \in A} \sum_{\substack{y \in A \\ y \neq x}} s^*(x,y)$$

$$= \frac{1}{2} \sum_{x \in A} \sum_{\substack{y \in A \\ y \neq x}} 1 \qquad \text{for maximum similarity}$$

$$= \frac{1}{2} N_A (N_A - 1)$$

Similarly

$$s^*(B,B) = \frac{1}{2} N_B(N_B - 1) \tag{10}$$

where $N_B = |B|$. Also

$$s^*(U,U) = \frac{1}{2} N(N - 1) \tag{11}$$

where $N = |U|$.

Let $s^*(A,B)$ be the value of $s(A,B)$ when all terms in A are maximally similar to all terms in B. Then:

$$s^*(A,B) = N_A \cdot N_B \tag{12}$$

since

$$s^*(A,B) = \sum_{x \in A} \sum_{y \in B} s^*(x,y)$$

$$= \sum_{x \in A} \sum_{y \in B} 1 \qquad \text{for maximum similarity}$$

$$= N_A \cdot N_B$$

It follows trivially that

$$s^*(A,A) + s^*(A,B) + s^*(B,B) = s^*(U,U) \tag{13}$$

For completeness:

$$N_A + N_B = N \tag{14}$$

The tendency of terms of A to be strongly similar to each other is indicated by

$$I_A = s^*(A,A) \; / \; s(A,A) \tag{15}$$

I_A decreases to one as the number and magnitude of the similarities between terms in A increases. The tendency for terms of A to be more similar to the terms in B than to other terms in A is indicated by

$$E_A = s(A,B) \; / \; s(A,A) \qquad (16)$$

E_A decreases to zero as the number and magnitude of the similarities of terms in A to terms in B decreases. I_B and E_B are defined analogously. The strength of the similarity between A and B compared with the maximum is given by

$$E_{AB} = s(A,B) \; / \; s^*(A,B) \qquad (17)$$

E_{AB} decreases as the number and magnitude of the similarities between A and B decrease.

The quantities I_A, I_B, E_A, E_B, and E_{AB} are independent of the scale of similarity, q, since all numerators and denominators are homogeneous and of the same degree in the term-term similarities s(x,y). Minimizing I_A corresponds, therefore, to maximizing the number and magnitude of the similarities between the terms of A. Minimizing E_A corresponds to minimizing the number and magnitude of the similarities between terms in A and terms discarded from A, namely those in B. Thus minimizing I_A and E_A jointly corresponds to constructing classes of terms which have the general character that terms belonging to a class are more strongly related to the other terms in the class than they are to terms not in the class. The universe of terms is therefore partitioned into two disjoint sets A and B by minimizing jointly I_A and E_A. The function of I_A and E_A which expresses the relationship between them as they are jointly minimized is not yet specified, but may be stated in the following general form:

$$c_A = f[s(A,A),\ s(A,B),\ s(B,B),\ N_A,\ N_B]$$

This function is called the <u>cohesion function</u>. c_A is the value of the cohesion across the boundary between the sets A and B. A set of terms which minimizes the cohesion is called a clump as defined by Parker-Rhodes and Needham [48].

Accordingly, the classification of the term vocabulary U into groups of similar terms will be defined in terms of minimum cohesion as follows. A class A is a clump provided that the cohesion for A is <u>locally minimum</u>. The cohesion for a class A of terms is locally minimum if the cohesion of each set in the <u>k locality</u> of A is not greater than the cohesion for A. A set P_i of terms is in the k locality of A provided that P and A differ in at most k terms, where k is a positive integer constant which may be chosen arbitrarily. Thus A is a clump if

$$c_A < c_{P_i} \qquad \text{for all } P_i \text{ such that } |P_i \wedge \bar{A}| + |\bar{P}_i \wedge A| < k$$

The classificatinn consists of the set of classes which satisfy this criterion. The membership of a term to a particular clump does not affect its membership to other clumps. Clumps may therefore overlap. However, the extent to which pairs of clumps may overlap,and the number of clumps in the classification is affected by the value of k. The number of local minima of the cohesion function, and therefore the number of clumps, cannot be determined since it depends upon the values of the individual similarities between pairs of terms. However, the following remarks may be made:

(1) Clumps may not differ in k terms or less. (For, if two sets differ in k terms, then each is in the k locality of the other. The set which has the smaller cohesion is a clump.)

(1') As a corollary: if there is a clump with k' terms, where $k' < k$, then there cannot be another clump with $k - k'$ terms or less, since if there were they would differ in k terms or less.

(2) The number of sets of terms in U which differ in more than k terms decreases with k. The number of clumps in U , therefore, decreases with k.

(3) The number of sets in the k locality of an arbitrary set increases with k.

For each putative clump A, all the sets in the k locality of A must be examined to establish that the cohesion for A is a local minimum. It follows from (3) that the amount of computation required to test that A is a clump, is least for classification based on $k = 1$, It follows from (1) that, for classifications based on $k = 1$, clumps must differ by at least two terms. The choice $k = 1$ therefore imposes least restriction on the structure of the classification, both in terms of the number of clumps which may be constructed, from (2), and in terms of the differences in composition of the clumps. For these reasons, the classifications are based on $k = 1$.

In the design of suitable cohesion functions, care must be exercised in selecting appropriate combinations of $s(A,A), s(A,B), s(B,B), N_A, N_B$ to minimize. The following rules[10] determine the general behavior which must be exhibited by cohesion functions. The quantities I_A, I_B, E_A, E_B, E_{AB} and their interpretation in terms of interset similarities have already been defined.

---

10. The definition of clumps in terms of local minima of cohesion has certain computational advantages. Rubin [54] constructs a set of rules for the design of "splitting functions" (analogous to cohesion functions) whose purpose is to define the "best" partition of a population into a number of disjoint sets. The search for the best partition progresses by maximizing the value of the partition function for each level of similarity. Rubin considers the danger in his algorithm of encountering local maxima which hinder the detection of the global maximum, and lists a number of heuristic techniques which help in this situation. The comparable difficulty with minima is not encountered in the construction of clumps, since clumps are defined in terms of local minima within a well-defined k locality.

(4) If the value of E_A for a putative clump A is increased by the removal of terms from A or by the addition of terms to A, then the cohesion must increase.

(5) If the value of I_A for a putative clump A is increased by the removal of terms from A or by the addition of terms to A, then the cohesion must increase.

(6) The cohesion must be independent of a change in scale of term-term similarities.

(7) A set, none of whose terms are similar to terms outside the set, must be a clump. In graph theoretic terms, the components of the undirected graph representing the term-term similarities must satisfy the criterion. Note that the components need not be maximally connected. This is achieved provided that

(7') The cohesion is always positive.

(7") If $s(A,B) = 0$ then $c_A = 0$.

(8) The cohesion function must have the property that $f[s(A,A),\ s(A,B),\ s(B,B),\ N_A,\ N_B]$ is either always greater than or always less than $f[s(A,A),\ s(A,B),\ s(B,B) + K_1,\ N_A,\ N_B + K_2]$ for each partition A,B of the term vocabulary, where K_1 and K_2 are positive or negative constants.

Requirement (8) is necessary as a consequence of the requirement that the addition of a set of documents, indexed only by a set of terms which do not belong to the vocabulary, should not disturb the classification of the vocabulary.

Suppose that the term vocabulary U of a document collection D is classified. A number of documents D' are appended to D to form a collection D" . The set of documents D' is indexed by a set of terms U' which includes none of U. The enlarged term vocabulary is denoted by U" . We require that a classification of U" should include all the clumps of a classification of U together with a classification of U' . None of the terms of U' may occur in the clumps of U since, on the basis of co-occurrence of terms in documents, no term in U is similar to any term in U' . Suppose that A is a clump belonging to a classification of U. The effect of adding the terms U' to U is to increase N_B by K_2 , where $K_2 = |U'|$ and to increase s(B,B) by K_1 , where $K_1 = s(U',U')$. s(A,A) and s(A,B) remain unchanged in value. Provided that requirement (8) holds, minimization will be unaffected so that A will be a clump in U" . Similarly, the removal (corresponding to $K_1, K_2 < 0$) of the set of terms U' from U" does not disturb the classification of the remaining terms of U" . Condition (8) is important computationally, for it implies that a set whose terms are similar to no terms outside the set may be excised from the term vocabulary and classified separately. The number of sets in the k locality of a putative clump in the excised set is smaller than the number of sets in the k locality of the putative clump considered as a clump of the complete term vocabulary. The amount of computation needed to establish that the class is indeed a clump is therefore reduced.

A number of cohesion functions have been designed which satisfy conditions (4) - (8). These are

$$g_1(A) = s(A,B) \;/\; [s(A,A) + s(B,B)] \tag{18}$$

Its intended action is to locate sets of terms A which are well separated from their complements. However, it has certain disagreeable properties. From equation (8), $g_1(A)$ may be reduced to

$$g_1(A) = s(A,B) \;/\; [S - s(A,B)]$$
$$= 1 \;/\; ([S/s(A,B)] - 1)$$

Minimizing $g_1(A)$ therefore corresponds to minimizing $s(A,B)$. There is a strong tendency for classification methods based on this function to obtain the null set as a clump, since there is nothing to prevent $s(A,B)$ from decreasing progressively, as terms are removed from A, except fortuitous similarities. A more workable function which has the same intended action is

$$g_1'(A) = s(A,B) \;/\; N_A \cdot N_B$$

that is,

$$g_1'(A) = E_{AB}$$

The tendency of $g_1'(A)$ to decrease as terms are removed from A is now counterbalanced by the factor $1/N_A$ which increases $g_1'(A)$. Both $g_1(A)$ and $g_1'(A)$ are symmetrical in A and B, so that if A is a clump then so is B.

$$g_2(A) = s^2(A,B) \;/\; s(A,A) \cdot s(B,B) \tag{19}$$

Its intended action is to separate the class A from its complement and to increase, at the same time, the number and magnitude of the similarities between terms in A. It may be expressed in the form:

$$g_2(A) = E_A \cdot E_B$$

The function is symmetrical in A and B. If A is a clump then the complement B of A is also a clump. The universe is therefore partitioned into two sets, both of which satisfy the criterion. The cohesion function has, in addition, the following property. Classifications are unchanged by a transformation of the term-set similarities of the form:

$$s'(x,A) = h(N_A)s(x,A)$$

where $h(N_A)$ is a function of N_A which is never zero, since:

$$g_2'(A) = s'^2(A,B) \;/\; s'(A,A) \cdot s'(B,B)$$

$$= [h^2(N_A)h^2(N_B)/h^2(N_A)h^2(N_B)][s^2(A,B)/s(A,A)s(B,B)]$$

$$= g_2(A) \; .$$

Classifications based on this cohesion function therefore seem to be singularly unaffected by space distortion.

$$g_3(A) = N_A(N_A - 1)[s(A,B) \;/\; s^2(A,A)] \tag{20}$$

This may be reduced to the form:

$$g_3(A) = E_A \circ I_A$$

This function is not symmetrical in A and B. It may be argued that it is overrestrictive to require that a necessary condition for a class to be a clump is that the complement should also be a clump. The action of $g_3(A)$ is similar to that of $g_2(A)$ apart from the symmetry requirement. A is constructed to be well-separated from its complement (measured by E_A) and to have a large number of strong similarities between its terms in proportion to its size (measured by I_A). However, classifications based on this function have tended to consist of large, heavily overlapping classes of loosely related terms. It appears therefore that the factor I_A has not been given sufficient emphasis during minimization.

$$g_4(A) = \frac{s(A,B)}{s(A,A)} \left\{ \frac{N_A(N_A - 1)}{2s(A,A)} - \frac{2p[s(A,A)]}{N_A(N_A - 1)} \right\} \tag{21}$$

This may be expressed in the following form:

$$g_4(A) = E_A[(I_A - p/I_A)]$$

where $0 \leq p \leq 1$.

Its action is to achieve a more flexible balance between the well separatedness of A from its complement B and the number and strength of the similarities between terms in A in proportion to its size. For small sets, the factor E_A dominates in value. As terms are added to A, the factor $(I_A - p/I_A)$ dominates in value. The minimum value of I_A is 1, and corresponds to complete and maximal similarity between all terms of A. As classes become more amorphous (i.e., fewer terms are similar and the similarities between the terms of A and itself become smaller) with the addition of terms from B, $(I_A - p/I_A)$ increases by an amount which is smaller the larger p is. In minimizing $g_4(A)$, terms in A are reallocated to reduce the magnitude of this factor and therefore to increase the strength and number of similarities between A and itself. The contributions of the two effects may be balanced by adjusting the value of p at the start of classification. Classifications are found to differ for $p = 0, \frac{1}{2}$, and 1 in a way predicted by the above analysis. Clumps for $p = 1$ tend, on the whole, to be better separated from their complements and to have stronger similarities between their members than for clumps based on $p = 0$. The effects of the values of p smaller than $\frac{1}{2}$ are negligible, and in the limit as p tends to zero, classifications are identified with those constructed with $g_3(A)$.

3.4 The Classification Algorithm

The strategy for detecting clumps of terms for an arbitrary cohesion function satisfying conditions (4) - (8) in Section 3.3 consists of three separate procedures, each of which will now be considered. The procedures are the starting procedure, dealing with the initiation of the search for a clump; the assignment procedure, dealing with the allocation and reallocation of terms to a putative clump; and the stopping procedure, dealing with the conditions under which the search for a clump may be terminated.

A clump has been defined as a set of terms for which the cohesion is locally minimum in the sense that, for all sets in the k locality of the clump, the clump has the least cohesion. For reasons given earlier, the clump finding strategy will deal only with the case $k = 1$. To find all of the clumps in a given population it is necessary to examine all the binary partitions (into sets A and B such that $A \cup B = U$ and $A \cap B = \emptyset$) of the population. A theoretically possible starting procedure, therefore, consists of enumerating all of these partitions. The clumps are a selection of those which satisfy the criterion of locally minimum cohesion. For a large population, however, complete examination of all binary partitions cannot be contemplated since the amount of computation required to do this is prohibitively large even if it is assumed that the components of the graph representing the term-pair similarities have been found.

One possible starting procedure consists of examining a subset of the total number of binary partitions of U into the sets A and B. Instead of testing whether A is a clump and discarding it if it is not, the composition of A is altered by the addition or removal of terms to improve the suitability of A as a clump. The terms which are added or removed are determined by the assignment procedure. Thus the testing of each partition for minimum local cohesion is replaced by the enumeration of a subset of the partitions of U into A and B and by the adjustment of A to conform with the criterion of minimum local cohesion. Since there is no a priori way of deciding whether a given partition eventually may be transformed into a clump by a series of adjustments, the choice of initial partition is arbitrary. Accordingly, the simplest initial partitions are examined, namely those for which $|A| = 1$. The single term which belongs to A is called the <u>starting element</u>. Thus clumps are constructed by accretion to a single term. The total number of initial partitions with $|A| = 1$ is $|U|$, the number of terms in the vocabulary. With such a starting procedure, it cannot be guaranteed that the total number of clumps present in the universe has been found. However, provided sufficient information about the relative similarities between terms may be recovered from the classification, we may assert that a sufficient number of clumps have been located. From a practical point of view, one is interested in constructing classes of terms for a specific application, and it is immaterial whether or not all of the classes have been discovered providing that the set actually found is adequate for the application.

Suppose that A is a putative clump and that t_i is a term. t_i belongs either to A or to B, the complement of A. For definiteness, suppose that t_i belongs to A. Suppose also that the cohesion for A is c_A. Consider then a new partition of U into the disjoint sets A', B' (i.e., $A' \cup B' = U$, $A' \cap B' = \emptyset$) formed by removing the term t_i from A and placing it in B. Thus $A' \hat{} t_i = A$ and $B \hat{} t_i = B'$. Suppose that $c_{A'}$ is the cohesion for A'. If $c_{A'} < c_A$ then the cohesion is reduced by the removal of t_i from A. Accordingly, A is a more suitable candidate as a dump than A. Suppose on the other hand, that $c_{A'} \geq c_A$, then the cohesion is increased by the removal of t_i from A'. A is therefore a more suitable candidate as a clump than A'. A comparable argument holds for the case where t_i belongs to B and is moved into A to form a new partition of U. The operation is repeated for the next term and one of A and A', whichever has the lower cohesion. The calculation of $c_{A'}$ necessitates the calculation of $s(A',A')$, $s(A' B')$, $s(B',B')$, $N_{A'}$, $N_{B'}$. It is unneccessary to reconsult the whole similarity matrix to do this, since the transition from the partition A,B to the partition A',B' involves the addition or removal of a single term t_i. Instead $s(A',A')$, $s(A',B')$, $s(B',B')$,

$N_{A'}$, $N_{B'}$ may be calculated quite straightforwardly from $s(A,A)$, $s(A,B)$, $s(B,B)$, N_A, N_B as follows. There are two cases.

(a) Suppose that t_i belongs to A. Then:

$$
\begin{aligned}
s(A',A') &= s(A,A) - s(t_i, A) \\
s(A',B') &= s(A,B) + s(t_i, A) - s(t_i,B) \\
s(B',B') &= s(B,B) + s(t_i, B) \\
N_{A'} &= N_A - 1 \\
N_{B'} &= N_B + 1
\end{aligned}
\tag{22}
$$

(b) Similarly, suppose that t_i does not belong to A. Then:

$$
\begin{aligned}
s(A',A') &= s(A,A) + s(t_i,A) \\
s(A',B') &= s(A,B) - s(t_i,A) + s(t_i,B) \\
s(B',B') &= s(B,B) - s(t_i,B) \\
N_{A'} &= N_A + 1 \\
N_{B'} &= N_B - 1
\end{aligned}
\tag{23}
$$

The values of $s(t_i,A)$ and $s(t_i,B)$ may be determined simultaneously by a single scan of the ith row of the similarity matrix. This technique may also be applied to the calculation of $s(A,A)$, $s(A,B)$, $s(B,B)$ for the starting partition A,B

where $|A| = 1$. Suppose that t is a starting element. Then:

$$
\begin{aligned}
s(A,A) &= 0 \\
s(A,B) &= s(t,B) \\
s(B,B) &= S - s(t,B)
\end{aligned}
\tag{24}
$$

where $S = s(U,U)$. The calculation of S involves a scan of the complete matrix. By using the above relations, a scan of the complete matrix once for each starting partition is avoided. During the course of classification, the matrix need be completely scanned only once.

Suppose that the term vocabulary U consists of the terms $t_1, t_2, \ldots, t_n$. Suppose also that A_0 is the putative clump. The assignment procedure constructs the sets $A_1, A_2, \ldots, A_{n+1}$ such that:

$$A_i \wedge t_{i+1} = A_{i+1} \qquad \text{(i.e, addition of } t_{i+1} \text{ to } A_i \text{ reduces cohesion)} \tag{25}$$

or

$$A_{i+1} \wedge t_{i+1} = A_i \qquad \text{(i.e., removal of } t_{i+1} \text{ from } A_i \text{ reduces cohesion.)}$$

or

$$A_{i+1} = A_i \qquad \text{(i.e, neither addition nor removal of } t_{i+1} \text{ reduces cohesion.)}$$

Suppose that m is the number of pairs of sets (A_i, A_{i+1}) such that $A_i \neq A_{i+1}$, that is, m is the number of terms which have been added to A_0 or removed from A_0 during a complete examination of U. As i goes from 1 to n , $A_{i+1} \wedge t_{i+1} = A_i$ and $A_i \wedge t_{i+1} = A_{i+1}$ generate all the sets A_{i+1} in the k locality of A_0 , where $k = 1$. Thus A_0 , by definition, is a clump provided $m = 0$, since no set in the k locality $(k = 1)$ of A_0 has smaller cohesion. If $m = 0$ the construction of A_0 is terminated, while if $m \neq 0$, A_{n+1} is offered as a more suitable candidate as a clump than A, and the procedure is repeated.

4. PSEUDO-CLASSIFICATION AND INFORMATION RETRIEVAL[11]

4.1 The Use of Relevance in an Evaluative Retrieval System

It seems evident from the work of Salton [56], and of Sparck Jones and Jackson [65] that a certain measure of improvement in performance[12] over simple term retrieval may be obtained by using classifications generated automatically. It has been established, independently, that small tightly structured classes of terms constitute a classification favorable to retrieval application, but it is as yet unknown whether the improvement gained by using such classifications may be increased still further. The classes used in these experiments have in the main been generated automatically by making use of the co-occurrence of terms in documents.

In the preceding sections, attention has been focused specifically on the effect of automatically generated classes on retrieval performance. Consideration has been given to both classification and information retrieval, and has been directed toward finding classification and retrieval algorithms which yield improvements in performance beyond that obtained by simple key-word coordination. An assessment of the effect

---

11. This material originally appeared in [25]. Permission to make use of it here has been granted by the American Society for Information Science.

12. Performance here relates only to the retrieval or non-retrieval of relevant or nonrelevant documents. No account is taken of "hardware factors." Thus, for example, the amount of effort expended in extracting relevant documents is not taken into consideration.

of two different classifications of the same document collection and the same set of requests on retrieval performance involves a comparison between the two recall-precision curves. This may be achieved by direct comparison of the recall-precision curves since the principal interest was in an overall improvement for all coordination levels. When the curves for two experiments crossed, improvement was regarded as uncertain. Salton and Lesk [55] have used statistical methods to establish whether a limited number of classifications and thesauri display consistent improvement over a number of different document collections. Other evaluative measures have been developed by Swets [70], who uses a decision theoretic approach, and by Romerio and Cavara [53].

The experimental studies outlined above are similar in a number of respects. First, they use the term descriptions of the documents in the collection to produce term co-occurrences, and then a classification of terms. Second, they attempt to evaluate the performance of the retrieval procedure by using the base set of requests and a table of documents relevant to each request. Third, there is no attempt to adjust a classification on the basis of information gathered during evaluation. This is not intended as a criticism of these approaches, for they set out to establish whether classifications derived solely from co-occurrences of terms within documents can improve performance. A subsequent adjustment to the classification is,therefore, outside the terms of reference of these investigations. Finally, none gives a clear indication of the best performance that theoretically may be obtained from the retrieval systems, given the document collection, the base set of requests, and the table of documents relevant to the base set of requests. Some work in this general direction, however, has been done by Cleverdon and Keen [10], who have examined the exhaustivity and specificity of several index languages.

There are four sources of information for an evaluative retrieval system, namely, the document collection, the term classification, the base set of requests, and the table of documents relevant to the base set of requests. The latter, together with a performance measure, enables an evaluation of the system to be made by comparing lists of documents actually retrieved with lists of documents which should be retrieved. Not all of the four sources of information are independent of each other. Under the hypothesis that:

(a) Co-occurence of terms within documents is a suitable measure of the similarity between terms

a classification may be generated automatically. This hypothesis will be referred to as the <u>Association Hypothesis</u>. Alternatively, the classification may be produced manually. The dependence between the document collection, the base set of requests, the term classification, and the table of relevant documents is, however, less tractable. Suppose, for example, that a match function, which measures the coefficient of matching between a document and a request, is designed, and that this coefficient varies with the coordination level of a match both on terms and on classes. Then with the hypothesis, which will be called the <u>Relevance Hypothesis</u>, that:

(b) Coordination is positively correlated with external relevance (i.e., that relevance may be defined algorithmically)

a classification and a match function could be imagined which retrieved all and only relevant documents, namely those with high match coefficients. In practice, however, this position is seldom attained; partly because the match function is an imperfect approximation to that function which actually corresponds to external relevance, if such exists; partly because the classification is defective; partly because there may, in fact, not be sufficient information provided to the system to achieve the best[13] performance; and partly because the notion of relevance may not be well formed and may result in inconsistencies of some sort. The difficulty is that the external judging of relevance is simulated by using an assortment of classification algorithms based on the association hypothesis and an assortment of match functions based on the relevance hypothesis. Moreover, the simulation is undirected since there is no guide as to how a better simulation may be achieved. Information has been injected into the system by introducing hypotheses, rather than by utilizing the information available in a more economical fashion. It is the utilization of the available information more economically and the ideas to which this leads which will now be examined.

---

13. Best performance is used here in the sense of complete agreement with the document-request relevance table.

4.2 The Use of Relevance in Pseudo-classifications

The relationships which hold between the categories of information used in an evaluative retrieval system have been described briefly. To make descriptions of this type more precise it is convenient to introduce the notion of a model. Suppose that a set of experiments is designed to examine a particular aspect of information retrieval. Functions and processes are designed according to a set of rules and these, when taken together, form the retrieval system. The set of rules which govern the construction of each part of the system represent a description of a particular approach to the solution of information retrieval. These rules themselves are hypotheses within the field of information retrieval and an evaluative retrieval system allows these hypotheses to be tested. A complete knowledge of the set of rules gives, ideally, complete knowledge of a particular approach to information retrieval. The set of rules is called a model and may be such that a number of quite different systems can be constructed to conform to them; each such system is called a representation of the model. Within a model of information retrieval there will be models, or submodels, for each of the logically distinct processes which together comprise the system. Thus, for example, there will be a classification model if classification is required as an adjunct of the retrieval system. There will be an information model describing the categories of information required by the system, together with statements of the assumptions made about these categories; a _relevance model_, containing statements about the way in which external relevance is simulated by the system; and finally, an _evaluation model_, containing statements of the assumptions made about the evaluation of the output of the retrieval system. Once the model has been defined, it may be found that experiments can be designed to test the validity of the complete model, that is, all representations within the model, rather than the validity of a single representation.

Section 4.1 contained an informal specification of part of a particular model of information retrieval in its reference to the relevance and association hypotheses [Section 4.1(a) and (b)] and the interrelation of the four categories of information listed in that section. In this model, the classification and match function are used together with the document descriptions to simulate the external judgments of relevance for the base set of requests. The information contained in the relevance table is disregarded in the construction of the system and is used _post facto_, evaluatively, to corroborate or contradict the result of the classification algorithm or the match function. Instead, suppose that each request is taken in turn and a classification of the terms is gradually constructed which confers a high match coefficient on those documents judged relevant and a low match coefficient on documents judged irrelevant. This presupposes that the requests are well-formed, and that the table of relevance appropriate to those requests is incontrovertible. The match function may be defective but it is assumed to be an approximation to an intuitive notion of relevance. Each constructive step consists of an operation altering the membership of selected terms to selected classes of the classification. Furthermore, a subsequent alteration of the classification to accommodate the relevance judgments of another request is prevented from affecting the classification in a way detrimental to the requests already examined. The grouping so constructed is a classification of the terms only since it consists, _de facto_, of classes of terms. There is little reason to assume that such a classification will consist of classes of terms which represent generic concepts associated with the document collection. Subsequent analysis, for example, by comparison with a classification generated from term co-occurrences, might demonstrate that this is so. In the absence of this, however, a distinction is drawn between these two classifications by referring to the former as a _pseudo-classification_, and it is to be regarded as a purely formal construction. Performance for the base set of requests will be, by construction, the best attainable for all representations of the classification model, given the particular model of retrieval.

5. THE USE OF PSEUDO-CLASSIFICATIONS IN RETRIEVAL

5.1 The Predictive Use of Pseudo-classifications

The procedure outlined here constructs limiting pseudo-classifications, which produce the best performance when used in conjunction with the apparatus of information retrieval. A distinction has been drawn between a pseudo-classification and a classification, for the former is a pure construction, while the latter is, as far as classification technique permits, a more fundamental grouping based on the resemblances between terms. The problem arises of predicting the effect of a pseudo-classification on retrieval performance when a new set of requests, distinct from the set used in constructing the pseudo-classification, is offered to the system. There is no guarantee that the system will respond in anything but a perverse way, since the pseudo-classification is derived from a particular set of requests and a particular set of relevance judgments over the whole document collection. It is hoped that the procedure has generalized the notion of relevance from a number of specific instances of relevance of documents to requests and has stored it in the pseudo-classification in such a way that the system may predict which documents bear the same relationship to subsequently presented requests as certain prescribed documents bear to the base set of requests; that is, by repeated analogical statements such as "the relationship between D_1 and R_1 is the same as the relationship between D_2 and R_2," where it is known that document D_1 is relevant to request R_1 and document D_2 is relevant to request R_2. Although generalization may take place, this cannot be assumed as the following hypothetical set of requests and relevance judgments demonstrates.

Suppose that each request is disjoint from all other requests, each request has one document relevant to it and no document is relevant to more than one request of the base set of requests. A possible pseudo-classification for this configuration is the assignment of each term to a class by itself and such a configuration does not extend to an arbitrary request. There are a number of reasons that indicate that this is unlikely, although experimental demonstration is required to show it in practice. The first point is that the match function is designed to reflect an intuitive notion of relevance formulated in the relevance hypothesis (b) in Section 4.1. The classification is constructed by operations involving the assignment of terms to existing classes and the creation of new classes by the assignment of terms not yet gathered together into a class. Pairs of terms within a class are used in retrieval as if they are <u>intersubstitutable</u> as defined in Section 2.1, and it is the intersubstitutable pairs, which yield good retrieval performance, that are to be located.

Although there may be some choice available in the selection of terms and classes to operate upon, and although these terms may be brought together into such classes randomly, in the event of no other basis for decision, it is reasonable to suppose that, as the classification develops it will become more determinate in that the opportunity for random assignments will diminish, since the number of assignments to be made is finite. Also, it is reasonable to suppose that the later assignments of terms to classes may have the effect of diminishing the deleterious effect of the random assignments made during the early stages of the construction of the classification provided that terms may be removed from classes. The problem of extending the system to deal with new requests will be referred to as the *problem of generalizing*.

The only satisfactory solution is to choose a base set of requests in such a way that the set is *representative* of all future requests to be submitted to the system. There are a number of ways of doing this. Suppose, for example, that a statement may be made about the distribution of terms in the requests that may be submitted to the system. Probabilities are assigned to the output of the system to indicate that, although a document may be deduced relevant to a request, it is only guaranteed relevant by the system with the specified probability. This is not the same as *degree of relevance* since the document may be entirely irrelevant to the request, and is unsatisfactory compared with the system's good response to requests of the base set.

Another possibility is that each request must satisfy prescribed conditions. The pseudo-classification is constructed only with requests satisfying these, and the system, in turn, is required to respond only to requests that are of this type. Such conditions may contain a statement about a classification of terms with reference to their co-occurrence within requests. Thus, while attempting to base a retrieval system on considerations that do not involve classification techniques *per se*, a classification is required in order to ensure that the procedure will generalize to subsequent requests.

Finally, there may be available a large collection of requests and the associated relevance judgments, and although it cannot be demonstrated categorically, they are highly likely to represent a typical and representative sample of all the requests that may be put to the system in the future, in that all the different kinds of requests are adequately represented. Although this remark is pragmatic, this state of affairs nevertheless may happen in a practical application.

5.2 The Evaluative Use of Pseudo-classifications

The use to which pseudo-classifications may be put in evaluating retrieval performance will now be considered by means of the following experiment. An information retrieval system is designed to conform to some model, and a match function is specified that measures the relevance of a document to a request. The design of this function is based on intuitive ideas about relevance and is an internal analog of relevance as judged by the users of the system. The function may be good or bad, according to its ability to retrieve relevant and only relevant documents, and may be subject to change and modification as ideas about the internal representation of relevance change. There are also a number of algorithms for producing term classifications conforming to a model of classification. It is necessary to isolate the classification, which, when used in conjunction with the match function, results in highest retrieval performance. There remains, however, the problem of evaluation. Although the use of precision-recall curves for this purpose seems to be well established, there is still criticism of them and there is no one method which is generally accepted. For example, Cleverdon and Keen [10] compute recall and precision without regard to the order in which documents are retrieved. Salton [56], on the other hand, is interested in determining whether the relevant documents are retrieved first. Suppose that a certain match function is decided upon for the experiment. Classifications are produced by a number of algorithms and among these the algorithm is to be found which yields the best retrieval performance, for a particular way of measuring performance. Accordingly, each of these classifications, in turn, is used in the retrieval system, and the resulting performances are compared to determine the best. These results may suggest a variation in classification technique and by experimentation of this kind it is hoped eventually to arrive at a classification algorithm that will give the best performance.

There is a serious difficulty with this approach. For a given document collection, base set of requests, and table of relevance there is a best possible level of performance, according to some measure of performance, the given match function, and the classification model. The measures of performance customarily used do not relate the performance of a retrieval system to the best theoretical performance for the retrieval model, with the result that there is no indication of the extent to which the system may be improved, or the direction in which such improvements may occur. Suppose, however, that it were possible to examine all the pseudo-classifications, in turn, which satisfied the classification model. The maximum theoretical performance for the model is then given by the classification which resulted in performance better than any of the other classifications, since all possible representations in the model would have been examined to make this assertion. Such a classification is the limiting pseudo-classification since it is the one that agrees most closely in retrieval with the relevance table for the base set of requests. Moreover, the pseudo-classification does not depend upon the measure of retrieval performance but more directly on the extent to which the relevance judgments set out in the relevance table have been reproduced by the system. The effect of the pseudo-classification may be assessed subsequently by choosing a particular measure of retrieval performance which uses this information, and the numerical values which result will represent the best possible performance obtainable with the given models.

The problem, however, cannot be approached in this way for, although the number of partitions, and, therefore, the number of classifications, of a finite number of terms is itself finite, the amount of computer time needed for evaluating them all is prohibitively large. Instead of enumerating them all, a method is needed for isolating those that are the best for the particular retrieval experiment. The extraction of classifications in the sense of linguistically coherent groups of terms is no longer of primary interest. Only formal groups of terms that result in good retrieval performance are of interest, because it is the hypothesis that a particular classification technique is the best for a specific retrieval experiment which is to be tested. These partitions are the pseudo-classifications, an algorithm for the construction of which has been alluded to earlier in Section 4.2.

The following way of measuring retrieval performance is suggested by the foregoing remarks. A particular match function and performance measure are chosen. A pseudo-classification is constructed (by algorithm rather than by selection from a complete enumeration of the representations of the classification model) which, when used in retrieval, gives rise to a particular level of performance. This is the theoretical best for the data, for the classification and retrieval models, and for the match function. The proposed measure of performance measures the departure of the performance in a particular case from this theoretical best.

If this departure is small, little improvement may be gained by changing the classification technique, indicating that further experimenting in this direction is unprofitable. The introduction of another match function, however, results in a different pseudo-classification and a different estimate of the theoretically best performance. If this is better than the optimum for the previous match function, then the choice of match function may be regarded as more useful in retrieval than the previous one in that it permits the system, at least theoretically, to achieve an improvement in performance. If the measure of retrieval performance is changed, it is not necessary to regenerate the pseudo-classification, since this does not depend on the formulation of the measure. The following definition is, therefore, made.

> The difference in performance in retrieval of the pseudo-classification and the automatically generated classification (according to some measure of performance) gives an estimate of the _improvability_ in performance of the retrieval model for the match function and classification algorithm which have been applied.

Although the use of this performance measure may indicate whether it is more profitable to improve performance by changing either the classification algorithm or the match function, no indication is given of how this is to be done. In spite of the fact that the classification of the model that gives the best performance is known, there is no means supplied for deriving it without recourse to the relevance judgments of the base set of requests. In spite of this, its usefulness as an evaluative device for experimental models is not impaired.

5.3 The Isolation of Inconsistencies in Relevance Judgments

The practicability of constructing a pseudo-classification must now be raised. It may happen that, after the classification has been constructed to give the correct response to a number of requests of the base set, a request is encountered whose processing conflicts diametrically with some of the conditions set up to prevent deterioration of the classification with respect to previously processed requests. It is possible that his results from an inconsistency in the judgment of relevance of documents to requests as supplied by experts in the field covered by the document collection. However, it is expected that the assessment of external relevance by an individual does not lead to serious inconsistency and that the same is true for the determination of relevance by consensus of opinion. Detailed work by Resnick and Savage [52] on the consistency of human judgments of relevance support this view. It may transpire that the number of conditions that have to be constructed to prevent deterioration, or the number of decisions that have to be taken and which are represented by the assignment of terms to classes, increases at such a rate with the number of requests processed that the classification becomes over-determined at an early stage of construction. The best that can be done with a request that leads to an inconsistency is to reject it and tolerate a small decrease in performance over the base set of requests. The origins of the inconsistencies may be difficult to locate, but at least it will be known that a particular document-request pair cannot be manipulated in the pseudo-classification to satisfy the conditions set up for earlier pairs without impairing retrieval performance. This itself serves as a guide to subsequent work on the construction of classification and match algorithms within the retrieval model.

On a purely practical level there is little defense against inconsistent judgments of external relevance. If relevance judgments were to be completely idiosyncratic, a retrieval system would have to be constructed for each user, and the only way forward would be by interactive techniques. If relevance judgments were to be arbitrary, no system could function at a predictable level of performance. A fixed, that is, noninteractive system relies on a consensus of opinion of users as to the relevance of documents to requests.

6. A RETRIEVAL MODEL

6.1. The Information Model

An informatinn model is a statement of the categories of information used to describe the system followed by a statement of the assumptions made about these categories. Within this model, algorithms may be designed to perform specified operations, and they must not refer directly or indirectly to information which is outside the model.

Section 4.1 referred to the four sources of information required by an evaluative retrieval system. These are

(1) D, the set of documents defined extensionally by terms. This is the document collection.

(2) R, the set of requests defined extensionally by terms. This is the base set of all requests.

(3) Z, the set of requests defined extensionally by the documents relevant to them. This is the set of relevance judgments.

(4) C, the set of classes defined extensionally by terms. This is the term classification.

These are similar in that each defines one set of elements extensionally in terms of another set of elements. Each, therefore, may be regarded as a rectangular incidence array giving the occurrence of an element of one type with an element of another type. The underlining indicates that we are considering the set of elements defined in terms of another set (i.e., an array) rather than a single element defined in terms of a set of elements. Attention is confined to documents indexed using simple term coordination, and document collections indexed probabilistically, for example, by the methods of Maron and Kuhns [38], are not considered. The reasons for this are threefold. First, the document collection used is not indexed probabilistically, although some attempt might be made to rectify this. Second, it is felt that tests with undifferentiated terms, logically, should precede any experimenting with term weighting in order first to establish a basis for comparison. Third, the construction of pseudo-classifications is simpler in this case since a single operation on the classification involves the dichotomous choice between the removal or insertion of a term. The probabilistic case is more complex since the choice of a value to be assigned to the weight is no longer dichotomous.

The terms which specify the requests also appear unweighted as do, for quite different reasons, the documents relevant to the base set of requests. It may be possible to construct a scale of relevance and assign degrees of relevance according to this scale. Instead, those documents have been selected that have been judged of highest relevance. Therefore, in this study, the scale of relevance as applied externally has two values, namely, "relevant" and "not relevant." It is realized that the external judgment of relevance is more complex than this and that there is no simple division of documents into those relevant to a request and those not relevant. The insertion of a third category would be more realistic, and such a category, namely "of unspecified relevance", would serve to separate the polarities of relevance more clearly. The quantification of the degree of relevance of documents to requests has further difficulties of quite a different nature. Suppose, for example, that it were possible to place in order of increasing relevance to a particular request all the documents of the collection. As soon as numerical values are assigned to each document to quantify its degree of relevance, metric properties are assumed about the scale of relevance. During the course of retrieval, arithmetic operations on degrees of relevance are performed which tacitly assume the truth of statements like, for example: "document D_1, whose degree of relevance to a request R is i, is i/j times as relevant to R as document D_2, whose degree of relevance to R is j." Until a metric has been established for degrees of relevance such statements remain indefensible. Resnick and Savage [52] have proposed to make a study of this question. For these reasons it is decided to work with categories of relevance rather than degrees of relevance; that is, with a qualitative scale rather than a quantitative scale. The two-valued scale has been adopted since the main body of data for testing purposes is reducible to this form.

6.2 The Classification Model

The model by which classifications to be constructed are constrained is as follows. Membership of terms to classes is a binary property; the object either belongs or does not belong to a class. The probabilistic assignment of terms to classes is excluded. All assignments of objects to classes are _a priori_ independent, and overlapping classes are allowed. Finally, the classification is nonhierarchical. It should be noted that the classes of the classification are defined extensionally. Terms are not assigned to classes according to their satisfying a known condition on the class. It may, nevertheless, transpire that classes have useful properties in terms of the character of the vocabulary.

The model of retrieval, therefore, is such that all sets encountered are defined extensionally and nonprobabilistically. For explanatory purposes we shall refer to Section 6.1 Statements (1) - (4) as the _retrieval environment_ (see (26)).

6.3 The Relevance Model

The match functions applied only make use of the information contained in or derivable from the environment. Thus:

$$\ell = M[D, R, f(D,\underline{C}), f(R,\underline{C})] \qquad (26)$$

where ℓ is the match coefficient corresponding to the match function M applied to an arbitrary document D in $\underline{D}$ and an arbitrary request R in $\underline{R}$ using the classification $\underline{C}$. f is a function which produces from a description of a set, specified using terms, a description using classes. The purpose of f is to provide a means of recovering in class matches the term matches that were missed on simple matching of the term descriptions for R and D because a term was used in one of these and a variant of this term in the other. The relationship between these terms may or may not be one of actual synonymy in natural language. The intention is that the classification should contain classes of terms mutually intersubstitutable and which result in good retrieval performance.

In accordance with the relevance hypothesis (Section 4.1), M is required to be a monotonically increasing function of the number of terms or classes in common between R and D. Its behavior with respect to the terms belonging to one of them, but not to the other, is not specified. This is a subject for precise formulation in a specific realization of M which satisfies the conditions mentioned. The modifications to the pseudo-classification as it is being constructed will be seen, from Section 7.3, to involve the addition of terms to classes already defined or the grouping of terms to form new classes. The size of the classes and their number will, therefore, vary during construction and no particularly relevant interpretation may be put on these quantities. In addition, if the match function is allowed to depend on them explicitly, the classification will be unalterable or will certainly deteriorate with respect to already processed requests, as new requests are examined. The dependence of the match function on these two quantities is, therefore, explicitly proscribed.

The values resulting from application of the match function to a document that is absolutely relevant to a request, and to a document that is absolutely irrelevant to a request, may be specified at will. Although absolute irrelevance is not as precise an intuitive concept as absolute relevance (for a reason can always be found for saying that a document is slightly relevant to a request in a collection of restricted subject matter), it will suffice simply to regard it as being that relationship which exists between document and request at the lowest value of the match coefficient. A corollary of the relevance hypothesis is that this coefficient must increase monotonically to its maximum value which represents absolute relevance. It is, therefore, required that the bounds of the match function be finite and that these bounds are attained, at least in theory, by the function. The additional conditions on the match function are, therefore:

(a) M is a monotonically increasing function of the number of terms or classes in common between R and D.

(b) M is independent of the size of any class in $\underline{C}$ and of the number of classes in $\underline{C}$.

(c) ℓ is bounded above and below, and attains its bounds.

Once the classification has been constructed using the base set of requests, it may still be underdetermined, but additional requests and their appropriate relevance judgments could be used to complete the classification only if the match function does not depend on the number of requests in the base set. Similarly, additional documents could only be added to the collection provided the match function does not depend on the size of the collection. Since both of these are valuable properties of a retrieval system, the following conditions on M are imposed:

(d) M does not depend on the size of R.

(e) M does not depend on the size of D, or on the size of the term vocabulary.

So far, there is no criterion of relevance in the model. Any criterion is bound to be arbitrary to a certain extent, for it is never possible to have complete knowledge, either of the document collection, or of the mind of the user of the system. To use the complete document unprocessed is as far from the solution as hoping to provide a complete analysis of the document, revealing in detail the complexity of the structural and semantic relationships between all the linguistic elements in the document. It is necessary to use approximations which hopefully will reveal the salient features of the collection for the purposes of automatic retrieval. In the model, the upper and lower bounds of the match coefficients that represent the polarities of relevance are known. By (26), a scale of relevance between these poles has been established and somewhere along this scale a value is defined, above which documents are retrieved and below which they are suppressed. This value is called the <u>critical value</u> of the match coefficient. Categories rather than degrees of relevance are assigned, therefore, to retrieved documents. In the absence of any evidence to the contrary, it will be assumed that the values of the match coefficient are distributed over the document collection in such a way that the critical value is the value midway between the extrema of the match coefficient. Subsequent experiment may require a revision of this assumption. Thus, for deciding on a suitable value for the critical point of the scale of relevance, the following hypothesis is made:

(f) The critical value is the arithmetic average of the upper and lower extrema of the match coefficient.

This will be referred to as the <u>critical value hypothesis</u>.

The following notation is introduced in connection with the above assumptions. Suppose that $\ell+$ and $\ell-$ are the upper and lower bounds of the match coefficient ℓ, respectively, and suppose that ℓ_0 is the critical value of ℓ. In accordance with the above hypothesis, ℓ_0 is defined as:

$$\ell_0 = \frac{1}{2}(\ell+ \ + \ \ell-) \ . \qquad (27)$$

The binary asymmetrical relations @ and $\not@$ are used to denote "relevant to" and "not relevent to," respectively. Thus:

$$\ell \geq \ell_0 \iff D \ @ \ R$$

$$(28)$$

$$\ell < \ell_0 \iff D \ \not@ \ R$$

For completeness, the binary asymmetrical relations $@^*$ and $\not@^*$ are introduced to denote "absolutely relevant to" and "absolutely irrelevant to," respectively. Thus:

$$\ell = \ell+ \iff D \ @^* \ R$$

$$(29)$$

7. METHODS FOR CONSTRUCTING PSEUDO-CLASSIFICATIONS

7.1. Two Approaches

Within the model of retrieval set out in Section 6, two approaches in the construction of pseudo-classifications may be distinguished. The first approach may be regarded as an attempt to an analytic solution by determining the "inverse" of the match function. This is seen more clearly from (26) which, in the context of Section 6.3, defines the match coefficient in terms of other elements of the model. This relationship is open to a more general interpretation in which it is regarded as an equation connecting ℓ, D, R, and $\underline{C}$ by the functions f and M, which are given. Providing only requests of the base set are considered, the values of ℓ, or at least their magnitudes with respect to the critical value of the match coefficient, are known. In the construction of pseudo-classifications, $\underline{C}$ is unknown so that (26) may now be interpreted as an implicit definition of $\underline{C}$. Such a solution, however, depends on whether M possesses the requisite algebraic properties which enable the inversion to be performed. Even if these properties were known, the method involves an amount of matrix manipulation -- for $\underline{D}$, $\underline{R}$ may be regarded as rectangular binary matrices -- which makes such a solution uneconomic for the scale of document collection eventually to be processed.

An alternative solution is by the method of perturbations, which has the advantage that less analysis of the match function is required. The classification is subjected to a number of alterations involving the insertion and deletion of terms from classes until the appropriate response to the base set of requests is elicited from the system. The alterations are carried out by perturbation functions.

7.2 General Outline of a Method

Suppose that it is possible, by a suitable perturbation, to cause the system to give the correct response, "relevant" or "not relevant", to a specific document-request pair. With a suitable convergence theorem, it would be possible to treat each document-request pair independently of the others. The response for one document-request pair might be destroyed or impaired by later adjustments to accommodate other pairs but at worst it would be necessary to examine each pair a number of times. The convergence theorem would ensure that although a number of responses may be impaired, gradual convergence would nevertheless set in. Without such a theorem, it is necessary to use a method in which each document-request pair is examined once only and in which conditions are set up to prevent a response from being obliterated by the processing of subsequent pairs. The degrading of responses by adjustment of the classification to accommodate subsequent document-request pairs is called deterioration. The conditions necessary to prevent the destruction of a response are referred to as deterioration conditions.

An outline of a method for constructing pseudo-classifications may now be given as follows:

(a) Let $\underline{C}'$ be the state of the pseudo-classification.

(b) Perturbations are applied to $\underline{C}'$ until the correct response is given by the system to a particular (D,R). The state of the pseudo-classification is then $\underline{C}''$.

(c) Deterioration conditions are set up for (D,R) with respect to $\underline{C}''$.

(d) The process is repeated for the next (D,R).

7.3 Design of Perturbation Functions

Deterioration conditions for the method described in Section 7.2 are effective only if all the perturbations which might lead to deterioration have been examined. It is, therefore, necessary that the perturbations which affect the match function are exhaustively enumerated and that there are few enough of them to make the method practicable. For each document-request pair, all perturbations which might lead to deterioration must be examined. It is required, as a practical constraint on the method, that the number of deteriorating perturbations be less than the number of document-request pairs to be considered. The algorithm for constructing pseudo-classifications is, therefore, of order two in the number of document-request pairs.

The operations which are appropriate to altering the classification are:

<u>Assignment</u> of terms to classes.

$$a(\{x_i\}^p \rightarrow \{y_i\}^p); \quad \langle \text{condition on} \{x\}^p, \{y\}^p \rangle \text{ for all } i \leq p \qquad (30)$$

<u>Removal</u> of terms from classes and their assignment to other classes.

$$r(\{x_i\}^p : \{y_i\}^p \rightarrow \{z_i\}^p); \langle \text{condition on } \{x\}^p, \{y\}^p, \{z\}^p \rangle \quad \text{for all } i \leq p \qquad (31)$$

where $\{x\}^p$ is an ordered set of p terms and $\{y\}^p$, $\{z\}^r$ are ordered sets of p classes, and where the conditions limit the choice of operands for a and r, x_i, y_i, z_i are the <u>i</u>th members of the sets $\{x\}^p$, $\{y\}^p$, $\{z\}^p$, respectively. p is called the <u>step length</u> of perturbatinn. A complete enumeration of the perturbations with <u>step length</u> <u>P</u> which affect the classification involves the enumeration of all perturbations of smaller step lengths (i.e., P - 1, P - 2,..., 2, 1).

Another possible perturbation is the simple removal of p terms from p classes. This, however, might result in the complete evacuation of terms from the classification. The removal of terms from classes is provided for in (31) and since, in this case, terms are reassigned to classes, there is no possibility of complete evacuation of the classification. For these reasons the simple removal of terms from classes without reassignment is not considered.

The effect of (30) on the classification is to classify terms (i.e, to insert terms into classes of the pseudo-classification) while the effect of (31) is to reclassify terms (i.e., to redistribute terms among the classes of the pseudo-classification). A further distinction is drawn. The classes mentioned in (30) and (31) are of two kinds. A class may already exist in the classification so that the effect of (30) and (31) is to add or remove terms. The effect, therefore, is to alter the constitution of a class. The number of classes in the classification remains unchanged or decreases. The classes may, however, be new in the sense that it is only when the functions have been applied that the classes enter the classification. They are not modifications of already existing classes since the number of classes in the classification increases. The effect of (30) and (31) when applied to such classes is to create new classes within the pseudo-classification. These distinctions are made use of in Section 7.5, where a method is proposed for selecting the appropriate perturbation to apply to a given document-request pair.

The choice of perturbation is further determined by the response which must be simulated for a given document-request pair. Suppose that the response of the retrieval system to the pair (D,R) is "not relevant" and that "relevant" is the correct response. A perturbation is applied to the classification which has the effect of increasing the value of the match coefficient for (D,R) to a level greater than the critical value. To facilitate this selection, the perturbations are grouped into those which increase the match coefficient, those which leave the match coefficient unchanged and those which decrease the match coefficient. The following terms are therefore used:

(a) _Increasing_ perturbations have the effect of increasing the match coefficient for a given document-request pair.

(b) _Level_ perturbations have the effect of leaving unchanged the match coefficient for a given document-request pair.

(c) _Decreasing_ perturbations have the effect of decreasing the match coefficient for a given document-request pair.

Once perturbation functions have been grouped according to their effects on the match function, it is clear how the deterioration conditions should be determined. If it is necessary to apply an increasing perturbation to accommodate a particular document-request pair, then conditions are set up to inhibit the application of the decreasing perturbations which may lower the match coefficient below the critical value. Similarly, if a decreasing perturbation is applied then conditions must be set up to inhibit the action of increasing perturbation functions.

The number of perturbations of the form (30) and (31) increases with the step length p. In order to satisfy the practical constraints on the model only single-step ($p = 1$) perturbatinns will be examined in detail.

7.4 Deterioration Conditions

Suppose that t_i is a term and that C_j is a class in the classification. It follows from the model of classification defined in Section 6.2 that either $t_i \varepsilon C_j$ or $t_i \notin C_j$ and that membership is nonprobabilistic. The classification $\underline{C}$ may, therefore, be represented by a binary array indicating the incidence of terms in classes. Thus,

$$C_{ij} = 1 \Longleftrightarrow t_i \varepsilon C_j \qquad (32)$$

$$C_{ij} = -1 \Longleftrightarrow t_i \notin C_j$$

where C is used to denote both the classification and the array. During the construction of the classification, however, another possibility of membership of terms to classes is used. The value $C_{ij} = 0$ is to imply that, on the basis of the information used so far, no decision may be taken about the membership of t_i to C_j although, at a later stage of construction, a definite choice may be made. During the construction of the classification the array is tri-valued. The values 1, 0, -1 are referred to as <u>status values</u>.

During the construction of the pseudo-classification, perturbations are applied which affect the membership of terms to classes and accordingly change the corresponding status values. A change in the status value is called a <u>transition</u> and is defined as::

$e_i \rightarrow e_j$ status value e_i is changed to status value e_j in the pseudo-classification. (33)

The transition table T_{ij} gives the set of permitted transitions and is defined by:

$$T_{ij} = 0 \iff e_i \rightarrow e_j \quad \text{not allowed}$$
$$T_{ij} = 1 \iff e_i \rightarrow e_j \quad \text{allowed} \tag{34}$$

where $e_i, e_j \varepsilon(-1, 0, 1)$. From (30) and (31) the single-step perturbations are of the form:

a(x → y); <condition on (x,y)> (35)

r(x:y → z); <condition on (x,y,z)> (36)

The deterioration condition associated with (35) is "x must never be assigned to y." This is achieved by forbidding the membership of term x to class y and by forbidding any transition from the status value -1. The effect on the pseudo-classification is, therefore:

$$C_{ij} = -1 \quad \text{where} \quad x = t_i \quad \text{and} \quad y = C_j$$

and the required values in the transition table to ensure that this is never revoked are

$$T_{-1.1} = 0 \quad \text{and} \quad T_{-1.0} = 0$$

The change to the classification is effected if

$$T_{0.\,-1} = 1$$

The deterioration condition associated with (36) is "if x is in y the x may not be assigned to z." This condition is recorded in the condition table Q(k:i,j) defined by:

$$
\begin{aligned}
Q(k:i,j) = 0 &\Longleftrightarrow r(t_k:C_i \rightarrow C_j) \quad \text{allowed} \\
Q(k:i,j) = 1 &\Longleftrightarrow r(t_k:C_i \rightarrow C_j) \quad \text{not allowed}
\end{aligned}
\qquad (37)
$$

These conditions may not be revoked in a later stage of the construction of the pseudo-classification. That is, within the condition table, the change from 1 to 0 is not allowed. The effect of (35) on the pseudo-classification is

$$C_{ij} = 1 \qquad \text{where} \quad x = t_i \quad \text{and} \quad y = C_j$$

and for this to be possible the transition $0 \rightarrow 1$ must be allowed:

$$T_{0.1} = 1$$

The effect of (36) on the pseudo-classification is

$$C_{ki} = 0, C_{kj} = 1 \qquad \text{where} \quad x = t_k, y = C_i, \ z = C_j$$

and for this to be possible the transition $1 \rightarrow 0$ must be allowed:

$$T_{1.0} = 1$$

The remaining transitions $e_i \rightarrow e_i$ for $e_i \in (-1, 0, 1)$ and $1 \rightarrow -1$ are allowed since they are not explicitly excluded by the above analysis. These results are collated bwlow. The transition table, therefore, may be indicated as

$$\begin{array}{cc|ccc|}
 & & & j & \\
 & & -1 & 0 & 1 \\
\hline
 & -1 & 1 & 1 & 1 \\
i & 0 & 0 & 1 & 1 \\
 & 1 & 0 & 1 & 1 \\
\hline
\end{array} \tag{38}$$

The effect of (35) on the classification array is

$$C_{ij} = 1 \tag{39}$$

The effect of (36) on the classification array is

$$C_{ki} = 0, \; C_{kj} = 1 \tag{40}$$

The deterioration condition for (35) is

$$C_{ij} = -1 \tag{41}$$

The deterioration condition for (36) is

$$Q(k{:}i,j) = 1 \tag{42}$$

Two functiona a' and r' are not introduced, which have the effect of setting up the deterioration conditions (41) and (42) associated with the functions a and r, respectively. The functions are called <u>conditional perturbations</u> and are defined as:

$$a'(x \rightarrow y); \qquad \langle \text{condition on } (x,y) \rangle \tag{43}$$

whose effect is $C_{ij} = -1$ where $x = t_i$ and $y = C_j$ [see (41)] .

$$r'(x:y \rightarrow z); \qquad \text{<condition on } (x,y,z)\text{>} \tag{44}$$

whose effect is $Q(k:i,j) = 1$ where $x = t_k$, $y = C_i$, $z = C_j$ [see (42)].

7.5 Precedence of Perturbations

It has been seen in Section 7.3 that perturbations may be grouped in two different ways: according to their effect on the match coefficient and according to their general effect on the classification. The selection of an appropriate perturbation to apply for a given document-request pair is determined, to a certain extent, by the system's response and the response specified in the relevance table. If, for example, the match coefficient for a document-request pair is lower than the critical value and it is known that the document is relevant to the request, then an increasing perturbation function is required. The choice of perturbation function is further determined by establishing a precedence between perturbations according to their effect on the pseudo-classification, namely classifying, reclassifying, or creating [see Section 7.3]. These three groups of perturbations will be denoted by c, r, and n, respectively. This is not, however, an exclusive grouping since a particular perturbation may be, for example, both r type and n type, that is, its effect is both to reclassify and to introduce a new class. A complete exclusive grouping consists of the group c(classify), r(reclassify), cn(classify and create new class), and rn(reclassify and create new class). The <u>type</u> of a perturbation is, therefore, defined as:

$$\text{type} = \text{c} \quad \text{or} \quad \text{r} \quad \text{or} \quad \text{cn} \quad \text{or} \quad \text{rn} \tag{45}$$

The first precedence to consider is that between r and c . During the construction of the pseudo-classification, it is advantageous first to attempt to reclassify the terms already classified until no further reclassification is possible. At this point more terms are admitted to the classification and are subsequently reclassified as appropriate. The precedence rule which achieves this effect is:

$$\text{r} > \text{c}$$

The same argument holds for rn and cn, to obtain the rule:

$$\text{rn} > \text{cn}$$

A final requirement is that the classification be constructed with as few classes as possible to achieve the required distinctions among the terms. This position is approached, at least in principle, if the perturbations involving the creation of new classes have the lowest precedence. Accordingly, the precedence of the types of perturbation is given by:

$$\text{precedence} = \text{r} > \text{c} > \text{rn} > \text{cn} \tag{46}$$

The selection of perturbations according to precedence and according to their effects on the match function does not lead to a unique function. There may be a number of perturbations which fulfill the conditions and for each of these there may be a number of possible choices of operands satisfying <condition on (x,y)> or <condition on (x,y,z)> [see (30) and (31)]. It will be seen that the perturbations change the match coefficient by an amount that is independent of the choice of arguments from among those which satisfy the appropriate conditions. Therefore:

(1) The arguments for the perturbation are chosen at random from the class of suitable operands.

To prevent the classificatinn from becoming overdetermined at an early stage of construction (in the sense described in Section 5.3):

(2) The smallest number of perturbations are selected which together produce the required change in the match coefficient.

CONCLUDING REMARKS

The analysis given provides the basis of a method for constructing the classification, that is the limiting classification, which enables the best retrieval performance to be achieved for a given set of relevance judgments and for a particular set of requests. Naturally, it is unrealistic to rely on the availability of relevance judgments, and the problem arises of constructing the limiting classification from the document collection alone. This problem remains unanswered at present. The most logical way in which to proceed is to examine the classes of the limiting classification to determine whether or not the members of each class are more strongly related to themselves than they are to terms outside the class. This question may be resolvable by comparing the limiting classification with classification constructed on the basis of the association hypothesis by techniques discussed in Section 3. Methods for comparing classifications are already available [24, 27]. Once the limiting classification is known, it is hoped that the methods outlined in this chapter will lead to the formulation of classification techniques which will be more applicable to the specific classification problems presented by information retrieval, than are those available at present.

ACKNOWLEDGMENTS

The author is indebted to Dr. R. M. Needham (University Mathematical Cambridge University, England) and Dr. K. Sparck Jones for their encouragement and advice throughout the course of the work presented here. He also wishes to thank Mr. D. S. Linney for his helpful criticisms.

This work is the result of research conducted on classification techniques for Information Retrieval supported in part by Grant Number GN-534.1 from the Office of Scientific Information Service of the National Science Foundation to the Computer and Information Science Research Center, The Ohio State University, and in part by a grant from the Office of Scientific and Technical Information of the United Kingdom.

REFERENCES

1. Adanson, M., *Famille des Plantes*, 2 vols., 1763. Vincent, Paris. Preface, III Partie, 1 er Article p. clv.
2. Ball, G. N., Data analysis in the Social Sciences: What about the details. *Proc. AFIPS*, Fall Joint Computer Conf., *Las Vegas*, 1965.
3. Benzecri, J.-P., *Sur les algorithmes de classification*. Texte multigraphie, 1965.
4. Benzécri, J.-P., Statistical analysis as a tool to make patterns emerge from data. *Methodologies Pattern Recognition*, (S. Watanabe, ed.) Academic Press, New York, pp. 35-74, 1969.
5. Bobrow, D. G., Natural language input for a computer problem solving system. Ph.D. Thesis, Mathematics Dept. M.I.T. Cambridge, Massachusetts, 1964.
6. Borko, H., Studies on the reliability and validity of factor analytically dervived classification categories. *Proc. Symp. Statistical Methods Mechanized Documentation, Washington*, 1964. Nat. Bur. Stds. Misc. Pubs., No. 269, pp. 245-251.
7. Borko, H., and Bernick, M., Automatic Document Classification. *J. Assoc. Comput. Mach.* 10, 151-162. (1963).
8. Bush, V., As we may think. *Atlantic Monthly*, 176 (1) July 1945, pp. 101-108. The Growth of Knowledge, in *Readings on Organization and Retrieval of Information* (M. Kochen, ed.), pp. 23-35. Wiley, New York, 1967.
9. Cleverdon, C. W., Report on effect of relevance decisions on comparative performance of index languages. College of Aeronautics, Cranfield, England, 1970.
10. Cleverdon, C. W., Mills, J., Keen, M., Factors determining the performance of indexing systems. Aslib Cranfield Research Project, 1966, Vol 1 (Parts 1 and 2) and Vol. 2. Cranfield, England.
11. Cooper, W. S. Fact retrieval and deductive question-answering information retrieval system. *J. Assoc. Comput. Mech.* 11, 117-137 (1964).
12. Cuadra, C. A. and Katter, R. V., Opening the black box of "relevance" *J. Doc.* 23, 291-303 (1967).

13. Darlington, J., Translating ordinary language into symbolic logic. Memo MAC-M-149, Project MAC, M.I.T., Cambridge, Massachusetts, 1964.
14. Dattola, R. L., A fast algorithm for automatic classification. *J. Libr. Automat.* 2, 31-48 (1969).
15. Diogenes Laertins, Lives of Eminent Philosophers. (R. D. Hicks, Transl.) Loeb Classical Library, 2 vols. W. Heinemann, London, 1925.
16. Doyle, L. B., Semantic road maps for literature searchers. *J. Assoc. Comput. Mach.* 8, 553-578 (1961).
17. Doyle, L. B., Breaking the cost-barrier in automatic classification. Systems Development Corporation paper SP-2516, 1966.
18. Doyle, L. B., Indexing and abstracting by association. *Am. Document.* 13, 378-390 (1962).
19. Doyle, L. B., Is relevance an adequate criterion in retrieval system evaluation? *American Documentation Institute: Automation and Scientific Communication. Ann. Meeting Amer. Doc. Inst., Chicago,* 1963, Part 2, p. 200. Amer. Doc. Inst., Washington, D.C., 1963.
20. Giuliano, V. E., Automatic message retrieval by associative techniques in *Joint Man-Computer Languages* Mitre Corporation SS-10, 1962. Bedford, Massachussetts, p. 10.
21. Gotlieb, C. C., and Kumar, S., Semantic clustering of index terms. *J. Assoc. Comput. Mach.* 15 493-513 (1968).
22. Green, B. F., Wolf, A. K., Chomsky, C., and Laughery, K. BASEBALL: an automatic question answerer, in *Computers and Thought.* E. A. Feigenbaum, and J. Feldman, eds., pp. 207-216. McGraw-Hill, New York, 1963.
23. Gull, C. D., Seven years of work on the organization of materials in the Special Library. *Am. Doc.* 7, 320-329 (1956).
24. Jackson, D. M., Comparsion of classifications. in *Numerical Taxonomy* A. J. Cole, ed. pp. 91-111. Academic Press, New York, 1969.
25. Jackson, D. M., Basis for an improvability measure for information retrieval. *Proc. 32nd Ann. Conv. Am. Soc. Information Science, San Francisco*, 1969, pp. 487-495.
26. Jackson, D. M., A note on a set of functions for information retrieval. *Inform. Star. Retrieval 5*,27-41 (1969).
27. Jackson, D. M., Automatic classification and information retrieval. Ph.D. Thesis, Faculty of Mathematics, Cambridge University, England, 1969.
28. Jardine, C. J., Jardine, N., and Sibson, R., The structure and construction of taxonomic hierarchies. *Math. Biosc.* 1, 173-179 (1967).
29. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* 12, 241-254 (1967).
30. Kuhns, J. L., The continuum of coefficients of association. *Proc. Symp. Statistical Methods Mechanized Documentation, Washington,* 1964. Nat. Bur. Stds. Misc. Publs., No. 269, pp. 33-39.

31. Lance, G. N., and Williams, W. T. A general theory of classificatory sorting strategies. II. Clurtening systems *Computer J.* 10, 271-277 (1967).
32. Lance, G. N., and Williams, W. T. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer J.* 9, 373-382 (1967).
33. Lewis, P. A. W., Baxendale, P. B., and Bennett, J. L., Statistical discrimination of the synonym/antonymy relationship between words. *J. Assoc. Comput. Mach.* 14, 20-44 (1967).
34. Library of Congress. Subject Headings, 6th Ed. Washington, D. C., 1957.
35. Luhn, H. P. A statistical approach to mechanised encoding and searching of literary information. IBM J. *Res. Devel.* 1, 309-317 (1957).
36. Mahalanobis, P. C., On the generalised distance in statistics. *Proc. Nat. Inst. Science India,* 2, 49-55 (1936).
37. Maron, M. E., Automatic indexing: an experimental enquiry. *J. Assoc. Comput. Mach.* 404-417 (1961).
38. Maron, M. E., and Kuhns, J. L., On relevance, probabilistic indexing and information retrieval. *J. Assoc. Comput. Mach.* 216-244 (1960).
39. Maron, M. E., Kuhns, J. L. and Ray, L. C., *Probabilistic Indexing.* Data Systems Project Office, Ramo-Wooldridge, Los Angeles, California. Tech. Memo. No. 3, 1959.
40. Mill, J. S., *System of Logic,* Vol. 1. J. W. Parker and Son, London, 1851.
41. Mooers, C. N., Summary of Lectures No. 1 and No. 2. Presented at the NATO *Advan. Study Inst. Automat. Doc. Analysis, Venice, July* 1963.
42. Needham, R. M., The place of automatic classiciation in information retrieval. Rept No. ML 166, Cambridge Language Research Unit, Cambridge, England, 1963.
43. Needham, R. M., Classification and grouping. Ph.D. Thesis. University of Cambridge, England, 1961.
44. Needham, R. M., The termination of certain iterative processes. Memo. No. RM-5188-PR. Rand Corporation, Santa Monica, California, 1966.
45. Needham, R. M., and Sparck Jones, K., Keywords and Clumps. *J. Doc.* 20, 5-15 (1964).
46. Parker-Rhodes, A. F., Statistical aspects of fungus forays. *Trans. Brit. Mycol. Soc.* 38, 283-290 (1955).
47. Parker-Rhodes, A. F., and Jackson, D. M., Automatic classification in the ecology of the high fungi, in *Numerical Taxonomy* (A. J. Cole, ed.), Academic Press, New York, 1969, pp. 181-215.
48. Parker-Rhodes, A. F., and Needham, R. M., The theory of clumps. Rept. ML 126, Cambridge Language Research Unit, Cambridge, England, 1960.
49. Pearson, K., On coefficients of racial likeness. *Biometrika* 18, 105-117 (1926).
50. Rao, C. R., *Advanced Statistical Methods in Biometric Research.* Wiley, New York, 1952.

51. Raphael, B., SIR-A computer program for semantic information retrieval. *Proc. AFIPS Fall Joint Computer Conf., San Francisco, 1964,* pp. 577-589.
52. Resnick, A. and Savage, T. E., The consistency of human judgements of relevance, *Am. Doc.* 15, 93-95 (1964).
53. Romerio, G. F., and Cavara, L., Assessment studies of documentation systems. *Inform. Stor. Retrieval* 4, 309-325 (1969).
54. Rubin, J., Optimum classification into groups: an approach for solving the taxonomy problem. Rept No. 39.014. IBM. New York Scientific Center, 1966.
55. Salton, G. and Lesk, M., The SMART automatic document retrieval system-an illustration. *Commun.* ACM 8. (6), 391-398 (1965).
56. Salton, G., Computer evaluation of indexing and text processing. *J. Assoc. Comput. Mach.* 15 (1), 3-36 (1968).
57. Salton, G. and Lesk, M. E., Relevance assessments and retrieval system evaluation. Inform. Stor. Retrieval, 4, 343-359, (1969).
58. Sebestyen, G.S. Recognition of membership in classes. IRE *Trans. Information Theory,* IT 44-50 (1961).
59. Shera, J. H. Libraries and the organization of knowledge. 1965 (D. J. Foskett, ed.). Archon Books, Hamden, Connecticut, 1965.
60. Simmons, R. F., Klein, S., and McConlogue, K. Towards the synthesis of human language behavior. Behavioral Science, 7, 402-407 (1962).
61. Sneath, P. H. A., Mathematics and classification from Adanson to the present in *The Bicentennial of Michel Adanson's "Famille des Plantes"* The Hunt Botanical Library, Carnegie Institute of Technology, Pittsburgh, Pennsylvania, 1964.
62. Soergel, D., Some remarks on information languages, their analysis and comparsion. *Inform. Star. Retrieval* 3, 219-291 (1967).
63. Sokal, R. R., and Sneath. P. H. A., *Principles of Numerical Taxonomy.* Freeman. San Francisco, and London 1963.
64. Sparck Jones K., Synonymy and semantic classification. Ph.D. Thesis, University of Cambridge, 1964.
65. Sparck Jones K. and Jackson, D. M., Current approaches to classification and clump-finding at the CLRU, *Computer J.* 10, 29-37 (1967).
66. Stevens, M. E., Automatic indexing: A state-of-the-art report. NBS Monograph No. 91. Natl. Bur. Stds. U.S. *Dept. Commerce,* U.S. Gov. Printing Office, Washington, D.C., 1965.
67. Stiles, H. E. The association factor in information retrieval. *J. Assoc. Comput. Mach.* 8. 271-279 (1961).
68. Sutcliffe, J. P. A probability model for errors of classification. I. General considerations. *Psychometrika* 30, 73-96 (1965).
69, Sutcliffe, J. P., A probability model for errors of classification. II. Particular cases. *Psychometrika* 30, 129-155 (1965).

70. Swets, J. A., *Effectiveness of Information Retrieval Methods*, Rept. 1499. Bolt Beranek and Newman, Cambridge, Massachusetts, April, 1967.
71. Tanimoto. T. T. An elementary theory of classification and prediction. IBM *Report*, Yorktown Heights, New York, 1958.
72. Taube, M. A note on the pseudo-mathematics of relevance. *Am. Doc.* 16, 69-72 (1965).
73. Vaswani, P. K. T. A technique for cluster emphasis and its application to automatic indexing, Booklet G. pp. 1-4 IFIP *Conf., Edinburgh, Scotland* North Holland Publishing Company, Amsterdam, Holland, 1968.
74. de la Vega, W. F. Techniques de classification automatique utilisant un indice de ressemblance. *Rev. Fr. Sociol.* 8, 506-520 (1967).
75. Verhoeff, J., Goffman. W. and Belzer, J. Inefficiency of the use of Boolean functions for information retrieval systems. *Commun.* ACM 4, 557-558. 594 (1961)
76. Williams, W. T., and Dale, M. B. Fundamental problems in numerical taxonomy. *Advan. Botan. Res.* 2, 35-68 (1965).