Information Technology: Research and Development (1982), *1* (157–171)

INFORMATION RETRIEVAL RESEARCH: STRATEGIES AND USER IMPLICATIONS*

V. V. RAGHAVAN

Computer Science, University of Regina, Regina, Canada, S4S 0A2

AND

J. S. DEOGUN

Computer Science, University of Nebraska, Lincoln, NE 68588, USA

(Received 20 August 1981, revised 11 November 1981)

ABSTRACT

Databases in common use are designed on the premise that the records can be described adequately by objective attributes alone. However, there are many situations where a need to use subjective attributes also exists. Information retrieval research deals with strategies for handling databases having attributes of both kinds. A model of the retrieval environment, titled the vector space model, which has been found to be extremely valuable in developing techniques to handle records represented by subjective attributes, is described. Some studies based on that model and their implications for the user are discussed.

1. INTRODUCTION

Information storage and retrieval is a discipline involved with the organization, structuring, retrieval and display of information. One of the important characteristics that distinguishes information retrieval from the well-known data management systems is that this discipline is concerned with *reference retrieval.* That is, the search result indicates where the answers to the user query may be found. In this sense, information retrieval systems deal with bibliographic databases, for example, databases consisting of books, reports, journal articles, etc. Information retrieval applications are, however, not limited to the library environment. A court file of previous cases and rulings, a government agency's file of patents, copyrights, etc.

* This research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

and a database of resource material in the research laboratories of large corporations are all examples of situations where the use of information retrieval systems can be beneficial.

One way of handling such bibliographic databases is by representing each item to be stored in the database by objective attributes (e.g., name of a product or publisher of a book) and by also ensuring that every attribute that might be of interest to the user is identified at an early stage of the design. In other words, the database must be first formatted. If this is done, commonly found database management systems can be used to provide access to and maintain the data. However, in many situations such as those mentioned above a need to (also) use *subjective* attributes or fuzzy descriptions (e.g., product has 'wood-like' texture, book is about 'thinking') exists. That is, the items that are to be dealt with are not amenable to precise description as is implied when only objective attributes are used. Consequently, special methods to deal with the problems and issues that arise in these situations must be developed. The research activities discussed in this paper pertain to such developments.

In order to put the nature of information retrieval in perspective, we draw an analogy between the requirements placed on management information systems and those of information retrieval systems. Burch *et al.* (1979) categorize decision situations into strategic, tactical or technical. Technical level decisions are usually routine and structured, and most low level management decisions (such as authorizing credit) are associated with this level. The routine data processing and record keeping activities are, usually, adequate to meet the requirements at this level. On the other hand, strategic level decisions are non-routine, ill-structured and require more intelligent effort, and decisions of this kind are generally associated with top level management. In this context, meeting the technical level information needs is easier than it is to support decisions at higher levels. As a result, the management information systems implemented in most companies never go beyond meeting the needs of the technical level. In comparison, the research efforts of interest in this paper are concerned with the development of techniques for automating processes (e.g., indexing, query modification) that are usually considered intellectual. Thus, these new techniques are to information retrieval as methods of meeting tactical and strategic level management needs are to management information systems. Interestingly enough, if the developments in information retrieval processes are adapted to data retrieval (as opposed to reference retrieval) applications, the resulting data management systems should be better able to meet higher level management needs. Perhaps, a manager might be able to ask which of his subordinates should receive a raise if the system he has allows him to specify, using subjective attributes, the factors to be considered in evaluating employee performance.

One of the objectives of this paper is to make the reader aware of the general approach to research, which has become somewhat of a standard over the years, in information retrieval. The *vector space model* (or vector processing model) proposed by Salton (1968, 1971) is used as a vehicle and the process of investigation is outlined by illustrating a situation where this model has been used to understand an important process in information retrieval. It is also our aim, in this paper, to survey important research findings that can, almost directly, be attributed to the use of this model. To make explicit the impact of research developments considered, the major features of existing retrieval systems and those of more recent experimental or prototype retrieval systems are discussed.

This paper is organized as outlined below. The next section outlines the important features of existing information retrieval systems. Section 3 presents the vector space model of the retrieval environment and discusses through an illustration how this model facilitates a better understanding of retrieval processes. A survey of recent research investigations is also presented. Section 4 outlines the characteristics of some experimental and non-commercial retrieval systems currently in use. These systems use, to varying degrees, the research findings presented in Section 3. The conclusions of the paper are then provided in the last section.

2. CHARACTERISTICS OF EXISTING SYSTEMS

Salton (1980a) identifies four main activities performed in an Information Retrieval System:

1. Information analysis
2. Information organization and search
3. Query formulation
4. Retrieval and dissemination.

The characteristics of existing systems are outlined in the light of this framework. The early systems concerned with information dissemination were off-line, batch systems. These were characterized by manually identified index terms (key-words), sequential file organization (using magnetic tapes), and Boolean queries; the search was done, for a number of accumulated queries, by making a single pass through the file.

There has been a major shift in the type of systems employed in the last several years in the sense that batch systems have been almost completely replaced by on-line systems. Most of the on-line systems currently in use employ an *inverted file organization.*

Query languages, however, have varied considerably in expressive power. Some systems merely provide searches based on (index) terms combined by the standard Boolean connectives whereas others allow the user to specify truncation on terms, provide weights of importance to terms and specify locational relationships between terms (e.g., appearance in the same sentence) (Burnaugh, 1967; Giering, 1972; IBM, 1978).

Due to the interactive nature of on-line systems, attention has been paid to the design of user interfaces that are more 'friendly'. Since many of the problems in the use of retrieval services have been caused by the vast differences in command languages and other facilities provided, there has also been an interest in developing a standardized interface that can be placed between the user and several retrieval services that he has access to. A typical interface of this kind is Conit, the Connector for Networked Information Transfer (Marcus and Reintjes, 1977). These developments should decrease the dependence of users on search intermediaries.

Most systems now provide displays of their search vocabulary and a list of synonyms or other terms that are related to the terms used in the user query. Another important feature provided in many systems is called 'browsing', whereby one or more previously retrieved documents are displayed. Clearly, these aids greatly enhance the query (re)formulation process.

There has also been considerable variation among systems from the point of view

of *information analysis.* The representation of items (documents) has varied 'from' the storing of the full text, 'through' a set of substrings or sentences selected from the full text, 'to' manually chosen (by subject experts) index terms. When index terms are used, they have normally been handled as if these attributes are objective (an item contains or does not contain a term). This process, of arriving at a representation for the items, is known as *indexing.*

Given these characteristics, the *retrieval* operation has involved the identification of the item lists corresponding to the terms in a query and the application of the Boolean operations to these lists as specified by the user. Depending on whether or not the query statement contained term weights, the output may or may not be ranked.

3. RESEARCH IN INFORMATION RETRIEVAL

Construction of abstract models of retrieval environment and/or processes has been an important aspect of progress made in this area. One of the most significant of these models is the vector space model which leads to effective and efficient processing of items characterized by subjective attributes. A number of studies based on this model have been reported by Salton (1968, 1971, 1975). Other models of retrieval processes have been based on fuzzy set theory, decision theory as well as considerations pertaining to probability theory (Tague, 1973; Bookstein and Swanson, 1974, 1975; Harter, 1975; Robertson and Sparck Jones, 1976; Tahani, 1976; Radecki, 1977; van Rijsbergen, 1977). Some studies have also appeared comparing the characteristics of these models (Robertson *et al.*, 1980; Salton, 1979; Yu *et al.*, 1979). Our aim, in this section, is to develop ideas relating to the vector space model and consider the impact that this model has had on information retrieval research.

3.1 The vector space model

Given a collection of documents, it is necessary to obtain a representation for the documents that allows the operations on the database to be carried out efficiently. In this context, it is assumed that each document is identified by a set of content descriptors (terms) extracted from one or more of the full text, the abstract, the title and so on. Suppose that the terms describing the various documents have been identified and that n distinct terms have been used in the index vocabulary. The collection may then be visualized as a matrix in which each row corresponds to a document and each column to one of the n terms. Alternatively, a document can be thought of as a vector in an n-dimensional space. The entry d_{ik} corresponding to the i^{th} document and the k^{th} index term may be binary valued, indicating just the presence or absence of the term in the document, or have numeric value reflecting the importance of that term as a descriptor of the document's content. A query Q is represented similarly by a term vector, $(q_1, q_2, \ldots q_n)$, where q_k indicates the value of the k^{th} term to the query.

For example, given two documents D_1 and D_2, both dealing with machines and thinking, D_1 may be described as (MACHINE,5;THINK,1) and D_2 as (MACHINE,2;THINK,4). We can, thus, distinguish D_1 (as dealing principally with machines) from D_2 (concerned more about thinking than about machines).

Similarly, a user interested primarily in documents about thinking may formulate his query as $Q = (\text{MACHINE},1;\text{THINK},6)$.

3.2 Search and retrieval operations

When documents and queries are represented by weighted vectors, unlike conventional retrieval systems where exact query–document comparisons are made, a more sophisticated system of query–document matching is adopted. The details of the matching process would vary depending on whether or not Boolean connectives are used in the query (Bookstein, 1980). But when a query is specified without interconnecting Boolean operators, one can choose from several vector matching functions. The simplest among these is the well-known inner product between vectors (also known as the *simple matching* function) given by

$$S_1(D_i,Q) = \sum_{k=1}^{n} d_{ik} \cdot q_k$$

When Q does not have any terms in common with D_i, the function has a value 0. As the match between the query and the document increases, the value of the expression increases. Thus, it is a similarity (rather than a distance) function. The upperbound of S_1 is however not well defined. This problem can be resolved, and the function made to have a value in the range 0 to 1, by adding a normalizing factor. An example of a function with this characteristic is as follows:

$$S_2(D_i,Q) = \frac{\sum_{k=1}^{n} d_{ik} \cdot q_k}{\sqrt{\sum_{k=1}^{n} (d_{ik})^2 \cdot \sum_{k=1}^{n} (q_k)^2}}$$

This function is referred to as the cosine correlation or *cosine similarity*, measures the cosine of the angle between the vectors D_i and Q in the n-space identified.

For the example previously introduced, the representations for D_1, D_2 and Q are $D_1 = (5,1)$, $D_2 = (2,4)$ and $Q = (1,6)$, since term 1 corresponds to MACHINE and term 2 to THINK. Thus, $S_1(D_1,Q) = 11$, $S_1(D_2,Q) = 26$, $S_2(D_1,Q) = 11/31$ and $S_2(D_2,Q) = 26/\sqrt{740}$. In both cases D_2 is determined to be closer to Q than is D_1. It is easily seen that these functions take into account not only the number of elements jointly present in two vectors, or jointly absent from both, but also the value of the elements concerned. There are a great many choices for finding similarity between vectors. It is also possible to use distance functions to determine the degree of match (or mismatch). Experiments with various functions, however, indicate that S_1 and S_2 are as effective in retrieval as other more complicated functions.

Once the representations and a similarity function are specified, the retrieval of documents can proceed as follows:

1. Determine the similarity between the documents and the query.
2. Rank the documents in descending order of similarity values.
3. Retrieve documents that have a sufficiently high rank.

The retrieval criterion may be specified either as a number that indicates how many documents the user wishes to see or as a threshold (cutoff) value to be applied to the query–document correlations. For our example, if function S_2 is chosen for retrieval and the threshold value chosen is 0.5, then only D_2 would be retrieved for query Q. Thus, document–query comparisons can be based on *best match*, rather than the exact match, criterion.

3.3 Performance evaluation

In order to develop retrieval systems that meet the user requirements satisfactorily, it is necessary to test numerous design alternatives in terms of information analysis, query (re)formulation, information organization, and so on. Although such testing should consider a full range of performance factors such as recall, precision, the effort involved from the point of view of the user, the response time, etc., the first two are the most critical in terms of the quality of response. Consequently, most research investigations have placed particular emphasis on improving performance as measured by recall and precision. These measures are given by:

$$\text{Recall} = \frac{\text{Number of documents relevant and retrieved}}{\text{Number of documents relevant}}$$

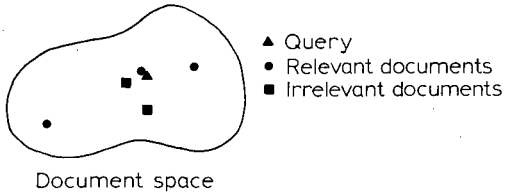
$$\text{Precision} = \frac{\text{Number of documents relevant and retrieved}}{\text{Number of documents retrieved}}$$

Note that a document is relevant or non-relevant in relation to a particular query and that this assessment is a function of user requirements. Furthermore, the computation of these values requires the specification of a threshold for retrieval in order that a distinction can be made between the retrieved and the non-retrieved documents.

Since the dependence on particular threshold and individual query is quite restricting, the evaluation employed in practice is made more elaborate. Typically, for a given document collection, a number of queries would be used during testing. For each query, the precision values would be computed at a number of discrete recall levels. That is, if the query has R relevant documents, the recall levels would correspond to retrieval thresholds being set at values where 1, 2, . . . , R relevant documents are retrieved (Fig. 1). The recall and precision values obtained for various queries can then be averaged to obtain a single set of recall and precision value pairs*. Those values can, of course, be plotted in a recall vs. precision graph. For an ideal system, where all the relevant documents obtain a ranking higher than any non-relevant document, the precision value would equal 1 for every recall value. More realistically, however, as higher recalls are desired, lower precision values would have to suffice and *vice versa.*

Consider, for instance, a situation where two methods of indexing (use of stems or thesaurus) are to be compared. Two retrieval runs can be made and recall and

* Many variations are possible on how these values might be summarized and reported. For details, see Salton 1971 and van Rijsbergen 1974.



Number of documents retrieved	*Recall*	*Precision*
1	1/3	1
4	2/3	2/4
5	1	3/5

FIG. 1. Changes in recall–precision values while number of documents retrieved increases. The distances between documents and query reflect their similarities.
▲ = query; • = relevant documents; ■ = irrelevant documents

precision values calculated as explained above. The averages for the two runs in the form of a graph is shown in Figure 2. The comparison can be made even more thorough through the use of statistical hypothesis testing methods (Salton and Lesk, 1968). Usually, the procedures outlined above would be repeated for a number of document collections.

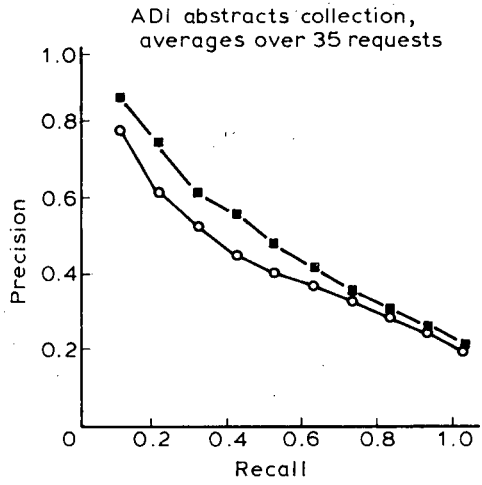


FIG. 2. Results of two retrieval runs, showing the averages produced.
o———o = stem; ■———■ = thesaurus
(Adapted from Salton, 1971, p. 80)

3.4 The use of the vector space model: an illustration

In this section, we consider a detailed example of how the vector space model enables a new understanding of the retrieval environment which, in turn, leads to methods of enhancing retrieval strategies.

The example draws from studies relating to automatic indexing. It is assumed that some kind of language processing techniques (see Salton, 1968, 1971, for details) have been used and, for each document in a collection, the index terms describing them have been identified. Often, the next step is to assign appropriate weights that characterize the usefulness of these terms and obtain further refinements to the

document representations. The *term discrimination model* (Bonwit and Aste-Tonsman, 1970) has been very important in the understanding of and development of methods for this latter step.

According to the term discrimination model, a good term is assumed to be one which, when used in the indexing vocabulary, will cause the greatest separation possible between the documents in the document space. On the other hand, a poor term is one which causes the documents to be more alike, and therefore makes it harder to distinguish one document from another. The idea is that the more dissimilar the document vectors are to one another, the easier it will be to retrieve some items while rejecting others; in contrast, when the documents are represented by similar term vectors it will be quite impossible to discriminate between the relevant and the non-relevant documents vis-à-vis a given query.

The usefulness of a term k is measured by its *discrimination value* (DV_k) defined as follows: let d_{ik} be as previously defined and let there be N documents in a collection. The centroid, or the centre of gravity, of all the document vectors can be written as a vector C where,

$$C = 1/N \sum_{i=1}^{N} d_{ik}$$

If the similarity between two vectors is measured by the function S_2, then the denseness or the compactness of the document space may be expressed as

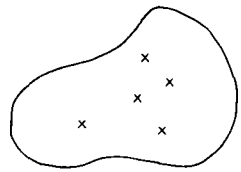
$$G = \sum_{i=1}^{N} S_2(C,D_i), 0 \leq G \leq N$$

that is, as the sum of the similarities of the centroid to the various documents. In terms of the above quantities, DV_k is defined as $G_k - G$, where G_k represents the denseness of the space with term k deleted from all document vectors. It is worth noting that the removal of term k has the effect of reducing the dimensionality of the space from n to $n-1$. It is easy to realize that $G_k > G$ for a good discriminator since the document vectors should be more bunched up in the space when the term is not used. Thus, for good terms $G_k - G > 0$. The opposite effect is observed for bad discriminators, so that $G_k - G < 0$. These concepts are graphically depicted in Figure 3.

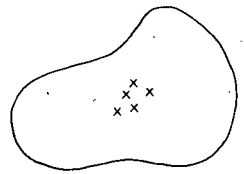
A straightforward way in which this analysis can be used in retrieval is to assign a weight of DV_k to term k. This approach, however, has not been found to give good performance. But improved indexing vocabularies could be constructed once the correspondence between DV_k and the number of documents of the collection in which term k appears (also known as *document frequency*) was established. Specifically, Salton *et al.* (1974) have determined by experiments that

1. the terms exhibiting the highest discrimination values (the good discriminators) are those with a medium value for the total number of occurrences in the documents (also known as *collection frequency*), and a document frequency less than one half its collection frequency;
2. next in discriminating power are the terms which have a discrimination value close to 0; these correspond to low document frequency terms;

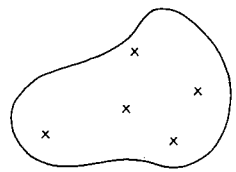
3. the terms which have negative discrimination values (the least attractive discriminators) are those that have a high document frequency (of the order of the collection size) and a collection frequency exceeding the collection size.



(a) Original document space



(b) Document space after removal of useful discriminators



(c) Document space after removal of useless nondiscriminators

FIG. 3. Changes in document space compactness following deletion of certain terms

These results led to the following strategy for improving indexing vocabularies: the terms in the low document frequency range must be combined into sets in such a way that the document frequencies of the resulting sets increase; contrariwise, the terms in the high frequency range must be broken up into subsets so as to produce terms with lower document frequency (Yu, 1973). These transformations correspond respectively to the processes of *thesaurus* construction (term classification) and *phrase* construction (identification of term phrases), which are well known in information retrieval.

The above example also brings out a characteristic very typical of information retrieval research, that a substantial amount of engineering work must be performed following the insight obtained through modeling and, possibly, analytical work.

Among the indexing methods known prior to that based on the term discrimination model, the most significant are due to Luhn (1957), Sparck Jones (1972) and Dennis (1967). It is interesting to note that Luhn was the first to conjecture that high document frequency terms and terms having very low document frequencies are

poor candidates for index terms, whereas the terms that are assigned neither to too many nor to too few documents are the best. The term discrimination model provides a clearer and more precise explanation of the relationship, between the document frequency of a term and its usefulness in retrieval, that Luhn suggested.

More recently, the understanding of the indexing process has been aided by the probabilistic models of the distribution of speciality words, proposed and tested by Bookstein and Swanson (1974, 1975), Harter (1975) and others. Another study that is of interest in this context is that by Yu *et al.* (in preparation) in which it is formally established that the nature of the relationship of document frequencies to term importance is in agreement with the suggestions of Luhn and the term discrimination model.

3.5 A brief review of related research

Advances have been made on the automation of several other (than automatic indexing) aspects of information retrieval system design. In particular, retrieval processes such as query reformulation, document clustering and construction of term classes are among the most frequently studied.

Query reformulation techniques address the problem of automatically modifying the user query based on feedback information obtained as to the relevance of previously retrieved documents. Rocchio and Salton proposed a technique that increases the weight of query terms contained in the documents judged as relevant and decreases the weight of query terms included in non-relevant items (Rocchio and Salton, 1965). This process is referred to as *relevance feedback*. It is expected that this process would obtain a query that is closer to the centroid of the relevant items retrieved than that of the non-relevant items retrieved. Assuming that the items relevant to the user are in close proximity to each other, the modified query should retrieve more of the relevant documents and fewer non-relevant ones. Although most of the studies in this area have been obtained in the context of queries represented as vectors of keywords, relevance feedback can also be adapted to retrieval environments that use Boolean queries (Vernimb, 1977).

A number of researchers have studied the problem of assigning weights of importance to query terms (Robertson and Sparck Jones, 1976; Yu and Salton, 1976; Robertson, 1977; van Rijsbergen, 1977). These represent an attempt to obtain a ranking of documents in the decreasing order of the probability of their being relevant to any given query under various assumptions. Since the approaches involved require relevance information, the weights so obtained have been referred to as *term relevance* or *term precision* weights. Thus, term relevance weighting methods represent a nice way of realizing the objectives of relevance feedback. Several attractive (automatic) query negotiation methods based on the findings of these studies have since been developed (Yu *et al.*, 1976, 1978; Harper and van Rijsbergen, 1978; Sparck Jones, 1979a, b; Robertson *et al.*, 1980; Chow and Yu, in preparation).

Document clustering is the process whereby documents in a collection are placed into affinity classes such that documents in the same class are more similar to each other than those from different classes. Given the constructs of the vector space model it is quite easy to visualize a document space consisting of such clusters. The investigations in this area have as primary objective the development of clustering algorithms that yield good clusters. Once the clusters are known, the search for

documents to be retrieved for a given query can be restricted to just the most promising clusters. Thus, organizing a file by clusters (*cluster file* organization) is much more natural in this environment than the traditional file organization techniques such as inverted lists, indexed sequential method, etc. Many recent studies have dealt with methods of document clustering (Salton, 1971; Croft, 1977, 1980; Salton and Wong, 1978; Dattola, 1969). Some of these investigations have shown that clustered file organization can not only make the retrieval process more efficient, but also more effective.

Construction of term classes, as suggested in Section 3.4, are important for building good term vocabularies. Furthermore, in systems where the user is given the option to modify the query manually, it is necessary to provide facilities for vocabulary displays that identify terms synonymous or otherwise-related to the terms included in the query. The most comprehensive investigation of the term clustering process, using as the basic parameter the frequencies of term pairs being assigned to the same documents (*co-occurrence frequencies*), has been by Sparck Jones (1971). An alternative approach to determining the relationship between terms which requires users' relevance judgments, called pseudo-classification, is introduced in (Jackson, 1970) and further developed in (Yu, 1975; Raghavan and Yu, 1979). A review of these and other studies on term classification can be found in (Salton, 1972, 1980b; Raghavan, 1978).

4. SOPHISTICATED RETRIEVAL SYSTEMS

A number of retrieval systems that tie in closely with the vector space model are now in use. On the one hand, these systems can be thought of as test beds for developing techniques that would enhance the performance of retrieval systems. That is, they provide an environment in which to perform research investigations of the kind discussed in the last section. On the other hand, a few of these systems are already being used in a real environment, and may be considered to represent the impact that the research investigations discussed earlier have had.

The SMART system (Salton, 1971), developed at Harvard and Cornell Universities, is the earliest retrieval system of this kind. The SMART system represents documents and queries as vectors, and allows best match retrieval to be performed. A number of basic routines are included for indexing, classification, relevance feedback and so on. The SMART routines can be run on the collections that are a part of its database, or on collections created by the user and added to the database using SMART facilities. Routines representing new methods of carrying out certain retrieval functions can also be added.

It should be noted that the collections included in the SMART system database are an important reason for the success of that system in its role as a test environment. A number of collections, in subjects such as documentation, aerodynamics, medicine and world events (based on news articles), are available. These collections have been prepared using the automatic language analysis techniques, developed by the SMART project workers, which take natural language (e.g., abstract of an article, a user's query stated in natural language) text as input and produce the weighted term vectors. Each collection consists not only of documents, but also of test queries made against the documents along with relevance judgments provided either by those who formulated the queries originally or by subject experts. Thus, the SMART system is an excellent research tool.

Another center where a great deal of research has been carried out in information retrieval is the Computer Laboratory at the University of Cambridge. Although there is little doubt that routines for many retrieval functions could be obtained from this laboratory, the authors are not aware of a retrieval system in operation. The Computer Laboratory can also provide access to many collections commonly used in research investigations.*

A number of retrieval systems have come up in the last few years that give one a flavor of the features that future systems are likely to have (Williams, 1969; Noreault *et al.*, 1977; Dattola, 1979; Landauer and Mah, 1979).

Browser (Williams, 1969) has facilities for processing natural language text somewhat like those of the SMART system. Each document (or query), which is in a natural language, is processed by the deletion of common (non-content) words and the words that remain are matched against a dictionary of manually prepared word-roots. The document number is then posted to every word-root recognized, obtaining an inverted file organization. The system also assigns a term weight (information value), to each word-root, that is inversely proportional to the number of occurrences of the word-root in the collection.

The METER (Landauer and Mah, 1979) system, which was originally intended for use in the analysis of current news items (e.g., investigative reporting), is an automatic information retrieval system, designed to exploit statistical associations between terms in documents consisting of English text. All phases of information analysis are automatic and the system provides an extremely simple user interface that is considered a substantial improvement over that of other similar systems. A 'scaled' version of METER that supports all major features of the full-scale METER system has been used for research and developmental work.

SIRE is a retrieval system with standard features (inverted file, Boolean queries, etc.). A recent study (Noreault *et al.*, 1977) shows that such a system can be enhanced by making minor modifications to the file design in a way that similarity functions can be used for matching purposes. As a result, ranked output in which relevant documents had a much higher ranking than non-relevant documents could be provided at a rather small additional cost. Strategies for ranking the output of Boolean searches in more complex situations are presented in Bookstein (1980).

FIRST (Dattola, 1979) is a document retrieval system which combines a database management system with automatic processing of natural language queries and document abstracts. In other words, this is a reference retrieval system that permits both objective attributes and subjective attributes to be handled, but each in a different way. Given a query having both kinds of attributes, the ranking procedure is similar to SIRE, except that the objective attributes (only) are first used to identify a subset of potentially relevant documents and, then, the documents in that subset are ranked on the basis of the subjective attributes they have in common with the query.

Crawford (1981) has investigated the possibility of adopting the relational view of data in the development of bibliographic retrieval systems. The modifications and extensions that would be required to the standard facilities provided by relational DBMSs are considered. The ideas developed in this work represent the basis of the MISTRAL project aimed at building a relational bibliographic retrieval system.

In the light of trends in information retrieval identified above, a summary of the

* A survey of various test collections that have been used in information retrieval research was carried out by Sparck Jones and van Rijsbergen (1976). The use of the same collections by most researchers has made findings in this area more easy to verify and interpret.

implications for future retrieval systems is presented in Table 1. In essence, the document and query representations would be obtained directly from natural language descriptions by using semi-automatic or fully automatic information analysis techniques. Clustered file organization would normally be used and the user would search the database interactively, through a friendly interface, with the aid of sophisticated facilities available for vocabulary displays, browsing, query reformulation and so on. The retrieval system would, by using close match criterion for retrieval, generate a ranking of the documents that reflects the weights of both query and document terms.

Table 1. A summary of trends in activities to be performed in retrieval systems

	Trends in processing techniques	
Activity	*From*	*To*
Mode	Batch, offline	Online
User interfaces	Aid user in query negotiation; require intermediaries	Friendly user interfaces with automatic query negotiation facilities
Query language	Boolean or other restricted language	Natural language
Information analysis	Full text or similar representation requiring minimal analysis or manual indexing by experts	Semi-automatic or fully automatic for processing natural language text
File organization	Inverted file	Clustered file
Retrieval criterion	Exact match	Best or close match
Output	Unranked, or ranked just using query term weights	Sophisticated ranking that considers both query and document term weights

5. CONCLUSION

The research strategies used in the development of reference retrieval systems are explained through a discussion of studies based on a model of retrieval environment called the vector space model. Most existing systems, while providing many sophisticated facilities for the user, are found to be lacking in their ability to handle subjective attributes. A look at the features of some test-bed type or experimental systems suggests that development of retrieval systems that would handle both subjective and objective attributes well is imminent.

REFERENCES

Bonwit, K. and Aste-Tonsman, J. (1970) *Negative dictionaries.* Cornell University: Scientific Report ISR-18, Section IV.

Bookstein, A. (1980) Fuzzy requests: an approach to weighted Boolean searches. *Journal of the American Society for Information Science 31*, 240–247.

Bookstein, A. and Swanson, D. R. (1974) Probabilistic models for automatic indexing. *Journal of the American Society for Information Science 25*, 312–319.

Bookstein, A. and Swanson, D. R. (1975) A decision theoretic foundation for indexing. *Journal of the American Society for Information Science 26*, 45–50.

Burch, Jr., J. G., Strater, F. R. and Grudnitski, G. (1979) *Information Systems: Theory and Practice.* New York: John Wiley & Sons, Inc.

Burnaugh, H. P. (1967) The BOLD (Bibliographic Online Display) system. *Information Retrieval—A Critical View.* (E. Schecter, ed.) pp. 53–66. Washington: Thomson Book Co.

Chow, D. and Yu, C. T. (in preparation) *On the construction of feedback queries.*

Crawford, R. G. (1981) The relational model in information retrieval. *Journal of the American Society for Information Science 32*, 51–64.

Croft, W. B. (1977) Clustering large files of documents using the single link method. *Journal of the American Society for Information Science 28*, 341–344.

Croft, W. B. (1980) A model of cluster searching based on classification. *Information Systems 5*, 341–344.

Dattola, R. T. (1969) A fast algorithm for automatic classification. *Journal of Library Automation 2*, 31–48.

Dattola, R. T. (1979) FIRST — Flexible information retrieval system for text. *Journal of the American Society for Information Science 30*, 9–14.

Dennis, S. F. (1967) The design and testing of a fully automatic indexing-searching system for documents consisting of expository text. *Information Retrieval — A Critical View.* (E. Schecter, ed.). Washington: Thomson Book Co.

Giering, R. H. (1972) *This is data central.* Dayton, Ohio: Mead Data Central Inc., Subsidiary of MEAD Corp. (Report DTN-72-2).

Harper, D. J. and van Rijsbergen, C. J. (1978) An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation 34*, 189–216.

Harter, S. P. (1975) A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science 26*, Part I: 197–205, Part II: 280–289.

IBM (1978) *Stairs VS — A Tool for the End User.* Uithoorn: IBM Scientific and Cross Industry Center, IBM Netherlands.

Jackson, D. M. (1970) The construction of retrieval experiments and pseudo-classification based on external relevance. *Information Storage and Retrieval 6*, 187–219.

Landauer, C. and Mah, C. (1979) Message extraction through estimated relevance. *Proceedings of the Second International ACM-SIGIR Conference.* pp. 64–70. Dallas, Texas.

Luhn, H. P. (1957) A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development 1*, 309–317.

Marcus, R. S. and Reintjes, J. F. (1977) Computer interfaces for user access to heterogeneous information retrieval systems. Cambridge, Massachusetts: Electronic Systems Laboratory, MIT (Report ESL-R-739).

Noreault, T., Koll, M. and McGill, M. (1977) Automatic ranked output from Boolean searches in SIRE. *Journal of the American Society for Information Science 28*, 333–339.

Radecki, T. (1977) Mathematical model of time-effective information retrieval system based on the theory of fuzzy sets. *Information Processing & Management 13*, 109–116.

Raghavan, V. V. (1978) *Evaluation of classification strategies for document retrieval*, Ph.D. Thesis, University of Alberta.

Raghavan, V. V. and Yu, C. T. (1979) Experiments on the determination of the relationships between terms. *ACM Transactions on Database Systems 4*, 240–260.

Robertson, S. E. (1977) The probability ranking principle in IR. *Journal of Documentation 33*, 294–304.

Robertson, S. E. and Sparck Jones, K. (1976) Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 129–146.

Robertson, S. E., van Rijsbergen, C. J. and Porter, M. F. (1980) Probabilistic models for indexing and searching. *Third International Conference on Information Storage and Retrieval.* Cambridge, England.

Rocchio, J. J. and Salton, G. (1965) Information optimization and interactive retrieval techniques. *Proceedings AFIPS, Fall Joint Computer Conference 27*, Part I, pp. 293–305. New York: Spartan Books.

Salton, G. (1968) *Automatic Information Organization and Retrieval.* New York: McGraw-Hill.

Salton, G. (Ed.) (1971) *The SMART Retrieval System — Experiments in Automatic Document Processing.* Englewood Cliffs, New Jersey: Prentice-Hall.

Salton, G. (1972) Experiments in automatic thesaurus construction for information retrieval. *Information Processing — 71.* pp. 115–123. Amsterdam: North Holland Publishing Co.

Salton, G. (1975) *Dynamic Information and Library Processing.* Englewood Cliffs, New Jersey: Prentice-Hall.

Salton, G. (1979) Mathematics and information retrieval. *Journal of Documentation 35,* 1–29.

Salton, G. (1980a) Automatic information retrieval. *IEEE Computer 13,* 41–56.

Salton, G. (1980b) Automatic term class construction using relevance — a summary of work in automatic pseudo-classification. *Information Processing & Management 16,* 1–15.

Salton, G. and Lesk, M. E. (1968) Computer evaluation of indexing and text processing. *Journal of the ACM 15,* 8–36.

Salton, G. and Wong, A. (1978) Generation and search of clustered files. *ACM Transactions on Database Systems 3,* 321–346.

Salton, G., Yang, C. S. and Yu, C. T. (1974) Contribution to the theory of indexing. *Information Processing — 74.* pp. 584–590. Amsterdam: North Holland Publishing Co.

Sparck Jones, K. (1971) *Automatic Keyword Classification for Information Retrieval.* Connecticut: Archon Books.

Sparck Jones, K. (1972) A statistical interpretation of term specificity in retrieval. *Journal of Documentation 28,* 11–21.

Sparck Jones, K. (1979a) Search term relevance weighting given little relevance information. *Journal of Documentation 35,* 39–48.

Sparck Jones, K. (1979b) Experiments in relevance weighting of search terms. *Information Processing & Management 15,* 133–144.

Sparck Jones, K. and van Rijsbergen, C. J. (1976) Information retrieval test collections. *Journal of Documentation 32,* 59–75.

Tague, J. M. (1973) A Bayesian approach to interactive retrieval. *Information Storage and Retrieval 9,* 129–142.

Tahani, V. (1976) A fuzzy model of document retrieval systems. *Information Processing & Management 12,* 177–187.

Van Rijsbergen, C. J. (1974) Foundation of evaluation. *Journal of Documentation 30,* 365–373.

Van Rijsbergen, C. J. (1977) A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation 33,* 106–119.

Vernimb, C. (1977) Automatic query adjustment in document retrieval. *Information Processing & Management 13,* 339–353.

Williams, J. H. Jr. (1969) *BROWSER: An automatic indexing on-line text retrieval system.* Gaithersburg, Maryland: IBM Federal Systems Division (Annual Progress Report — AD 693143).

Yu, C. T. (1973) *Theory of indexing and classification,* Report 73-181, Ph.D. Thesis. Cornell University.

Yu, C. T. (1975) A formal construction of term classes. *Journal of the ACM 22,* 17–37.

Yu, C. T., Lam, K. and Salton, G. (in preparation) Term weighting in information retrieval using the term precision model.

Yu, C. T., Luk, W. S. and Cheung, T. Y. (1976) A statistical model for relevance feedback in information retrieval. *Journal of the ACM 23,* 273–276.

Yu, C. T., Luk, W. S. and Siu, M. K. (1979) On models of information retrieval processes. *Information Systems 4,* 205–218.

Yu, C. T. and Salton, G. (1976) Precision weighting — an effective automatic indexing method. *Journal of the ACM 23,* 76–88.

Yu, C. T., Salton, G. and Siu, M. K. (1978) Effective automatic indexing using term addition and deletion. *Journal of the ACM 25,* 210–225.