

The Smart environment for retrieval system evaluation—advantages and problem areas

Gerard Salton*

The Smart environment provides a test-bed for implementing and evaluating a large number of different automatic search and retrieval processes. In this chapter, the basic parameters underlying the Smart system design are briefly outlined, and a comparison is made with the characteristics of more conventional retrieval systems. The principal lessons learned from the Smart experiments are described, and some of the methodological problems raised by the system design are outlined. Finally, some comments are included about the disadvantages inherent in working in the laboratory, and the insights that can be gained in such a situation.

15.1 Retrieval system environment

Automatic, or semi-automatic information search and retrieval systems have now been in existence for some twenty years. In the early years, only small collections could be searched, and the search requests received from the user population would be accumulated for some period of time, or 'batched' before actually being processed, with the result that several weeks would normally elapse before answers could be obtained to a given query.

At the present time, the role and importance of information retrieval has greatly increased for two main reasons: the coverage of the searchable collections is now extensive and collection sizes may exceed several million documents; furthermore, the search results can now be obtained more or less instantaneously, using online procedures and computer terminal devices that provide interaction and communication between system and users. The large collection sizes make it plausible to the users that relevant information will in fact be retrieved as a result of a search operation, and the probability of obtaining the search output without delay creates a substantial user demand for the retrieval services. It is not surprising in these circumstances that several million search requests are currently submitted each year to a variety of automatic retrieval services.

* This study was supported in part by the National Science Foundation under grant DSI-77-04843.

While the operational retrieval environment has thus drastically changed over the last few years, the intellectual design of the retrieval operations has remained reasonably unchanged for some decades. The following principal characteristics may be noted:

- (a) documents are normally indexed manually, that is, subject indicators and content descriptions are manually assigned to the bibliographic items by subject experts and professional indexers;
- (b) search statements are manually formulated by users or search intermediaries using one or more acceptable search terms and appropriate boolean connectives between the terms; subsequent reformulations and improvements in the query formulations are also carried out manually;
- (c) the principal file search device is an auxiliary, so-called *inverted* directory which contains for each accepted content descriptor a list of the document references to which that term is assigned; the documents to be retrieved are then identified by comparing and merging the document reference lists corresponding to the various query terms;
- (d) an 'exact match' retrieval strategy is carried out by retrieving all items whose content description exactly matches the term combination specified in the search request; normally, all retrieved items are considered by the system as being equally relevant to the user's needs, and no special method is provided for ranking the output items in presumed order of goodness for the user.

Enhancements are included in many of the modern search systems in the form of 'free text' manipulations allowing the user to choose arbitrary search terms, that is natural language terms that are not controlled by any dictionary or authority lists, leading to the retrieval of all documents whose stored texts (or text excerpts) contain a particular term combination included in the search requests. But even in the free text search mode, inverted directories are created containing all the text words that could lead to the retrieval of a given document in the collection. Additional refinements in the search mode are available in some modern online environments in the form of dictionary and vocabulary displays leading to better query formulation capabilities. However, the basic manual query formulation and exact match retrieval strategy based on inverted files is maintained in practically all operational retrieval situations.

When the work on the Smart retrieval experiments was initiated in the early 1960s, some attempts had been made at implementing so-called automatic indexing systems¹⁻⁴. These consisted in using the computer to scan document texts, or text excerpts such as document abstracts, and in assigning as content descriptors words that occurred sufficiently frequently in a given text. The early retrieval experiments conducted with such automatic indexing products showed that a large number of the automatically chosen index terms would also have been assigned by manual indexers, and that the automatic indexing products contrary to expectation did not prove to be totally inadequate.

Moreover, it appeared that the rudimentary early automatic indexing products could be easily improved. Thus linguists led the way by pointing out that a number of linguistic processes were 'essential' for the generation of

effective content identifiers characterizing natural language texts. Among the linguistic techniques of interest, the following were considered to be of greatest importance:

- (a) The use of **hierarchical term arrangements**, relating the content terms in a given subject area. With such preconstructed term hierarchies, the standard content descriptions can be 'expanded' by adding hierarchically superior (more general) terms as well as hierarchically inferior (more specific) terms to a given content description.
- (b) The use of synonym dictionaries, or **thesauri**, in which each term is included in a class of synonymous, or related terms. Using a thesaurus each originally available term can be replaced by a complete class of related terms thereby broadening the original context description.
- (c) The utilization of **syntactic analysis** systems capable of specifying syntactic roles for each term and of forming complex content descriptions consisting of term phrases and large syntactic units. A syntactic analysis scheme makes it possible to supply specific content identifications and avoids confusion between composite terms such as 'blind Venetian' and 'Venetian blind'.
- (d) The use of **semantic analysis** systems in which the syntactic units are supplemented by semantic roles attached to the entities making up a given content description. Semantic analysis systems utilize various kinds of knowledge extraneous to the documents, often specified by preconstructed 'semantic graphs' and other related constructs.

The design of the original Smart system was then based on the premise that effective automatic indexing procedures could be built by incorporating into a content analysis system one or more of the foregoing language processing methods. Most of the required constructs such as the hierarchical term arrangements and the syntactically analysed text excerpts could be represented by abstract trees, and other constructs such as semantic graphs and thesauri are easily represented by graph structures. Well known automatic procedures were also available for traversing and manipulating tree and graph structures⁵. The original Smart system was then designed to process natural language texts using these complex data structures.

To validate the linguistic analysis procedures it was necessary to compare the search results obtained by using term hierarchies and thesauri with other simpler systems based on the use of single, frequency-weighted terms extracted from the document texts. From the beginning, the Smart system thus contained an evaluation package based on the use of sample document and query collections and on the availability of full relevance assessments specifying the presumed relevance of each document with respect to each user query. This made it possible to compute for each processed query the **recall** and **precision** values measuring respectively the proportion of relevant items retrieved and the proportion of retrieved items that are relevant.

The early tests in turn led to additional experiments and to the development of a full evaluation system for a large variety of search and retrieval procedures. These developments are described in more detail in the remainder of this study.

15.2 Basic Smart system assumptions and early results

In the Smart system each record, or document, is represented by a **vector** of terms, that is $D_i = (d_{i1}, d_{i2}, \dots, d_{it})$, where d_{ij} represents the weight or importance of term j for document D_i . By 'term' is meant some form of content identifier such as a word extracted from a document text, a word phrase, a thesaurus class, an entry from a term hierarchy, etc. A query Q_j can be similarly represented as $Q_j = (q_{j1}, q_{j2}, \dots, q_{jt})$, and retrieval of a stored item can be made to depend on the magnitude of a global similarity coefficient $s(D_i, Q_j)$. Specifically, whenever $s(D_i, Q_j) \geq T$ for some threshold T , D_i is retrieved in answer to Q_j . It should be noted that an exact match between any particular query and document terms is never required for retrieval of an item. Instead, the similarity measure s may be based on the composite similarities between the full query and document vectors. Furthermore, since $s(D_i, Q_j)$ represents a measure of closeness between D_i and Q_j , the output documents can be presented to the user population in ranked order of presumed relevance to the user, that is, in decreasing order of the corresponding s coefficients.

The following assumptions are immediately implied by the vector processing environment:

- (a) In principle, each term included in a given vector is as important as any other term (except for the possible distinction implied by a particular term weight assignment); that is, each term represents a particular dimension in the t -dimensional vector space defined by the t terms used to index the document collection.
- (b) No relationships are defined between distinct terms; that is, the coordinate axes representing the distinct terms are assumed to be orthogonal.
- (c) A document is represented by a particular position, and possibly by a given length, in the t -dimensional vector space. (In practice, it is often convenient to normalize all vectors to some given standard length.)

In examining the Smart system, it is necessary to consider also another principal characteristic of the experimental environment, namely the use of small sample collections of documents and user queries for test purposes. Such a test environment makes it possible to carry out many different experiments at reasonable cost. Furthermore, a great many inconveniences inherent in the use of large operational collections are immediately eliminated. Thus full relevance assessments can be obtained from the user population of each document with respect to each query, leading to the generation of accurate recall-precision measures. The alternative would consist in using sampling techniques and obtaining relevance assessments for a portion of the document collection only. The use of sampling methods, however, introduces additional variables and the evaluation results may then be subject to substantial fluctuations.

The small document environment used in the Smart experiments also renders unnecessary the choice of various parameter values which would otherwise be required to control the retrieval process. Because the documents are ranked at the output in decreasing order of query-document similarity,

there is thus no need to choose a retrieval threshold to distinguish the retrieved from the non-retrieved items. Instead, recall-precision values can be computed for all possible retrieval thresholds—that is, after retrieving one, two, and eventually n documents in decreasing order of the similarity with the query—and the results can be plotted in a composite recall-precision graph. The experiments can then be carried out using a very small number of variable parameters such as collection size, number of queries, relevance assessments of documents with respect to queries, interpolation procedures for calculating precision values at fixed recall intervals, and methods for averaging the results over a number of different user queries⁶. The Smart experiments have thus come close to achieving the conditions often assumed for ideal retrieval test environments⁷⁻⁹.

The artificial collection environment does, however, have implications about the conclusions derivable from the experiments. Thus it is difficult to obtain really believable **efficiency** (as opposed to effectiveness) criteria, such as response time, processing cost, and user effort needed to submit queries and to obtain results, because no obvious procedure is available for extrapolating these efficiency measures to large, operational retrieval situations. Furthermore, when a restricted number of user queries is used to evaluate retrieval effectiveness, the implicit assumption is that these queries and the corresponding users are representative of a general user population at large.

For the Smart experiments, no attempts were made to generate efficiency data, and the requirements for a representative user population were met by extending the experiments to many different collections in different subject areas, and using many kinds of user queries. When two given processing methods are compared and the retrieval results for several different collections in distinct subject areas indicate that method A furnishes better retrieval output than method B, the indications are that these results reflect real differences in retrieval effectiveness. The repetition of a given experiment using several different test collections may also be useful in overcoming some of the sampling problems which arise when test collections with satisfactory statistical properties must be chosen. Furthermore, when a number of parallel results are obtained with different collections, the *relative* performance of the various processing methods may be measurable reasonably securely. *Absolute* performance values, on the other hand, are always difficult to use and interpret. Thus a precision performance of 0.20, indicating that one out of five retrieved documents appears relevant to the user's interests, may be acceptable when the recall is high and the number of retrieved documents is small; on the other hand, a larger precision of 0.50 may prove unsatisfactory in practice when the number of retrieved documents becomes too large or the recall is too low.

The first test results obtained with the Smart system in 1964 and early 1965 proved to be quite different from what had been expected. Invariably they showed that the more complicated linguistic methodologies which were believed essential to attain reasonable retrieval effectiveness were not useful in raising performance. In particular, the use of syntactic analysis procedures to construct syntactic content phrases, and the utilization of concept hierarchies could not be proven effective under any circumstances. The most helpful content analysis process seemed to be the extraction of weighted

word stems from document titles and abstracts, possibly supplemented by the use of a term classification, or thesaurus, designed to recognize some synonyms and related terms¹⁰⁻¹².

At first, the evaluation results were thought to be indicative of flaws in the system design, and the decision was made to redesign the Smart environment so as to create a more flexible retrieval environment. In time, several other large-scale retrieval tests carried out independently of the Smart environment have, however, confirmed the original Smart results. In particular, the well-known Aslib Cranfield project also found that the simpler indexing methodologies were more effective than the more complex ones, and at the present time, there is an understanding among retrieval experts that an overspecification of document content normally produced by the more refined indexing methodologies can be just as detrimental as an underspecification¹³. This evidence does not, however, prevent many people from still clamouring for more sophisticated linguistic analysis procedures to be incorporated into automatic indexing systems, or indeed from incorporating such methodologies into newly designed retrieval systems¹⁴.

The extended Smart system is briefly described in the next section and the various insights gained from the Smart experimentation are discussed in the remainder of the study.

15.3 The extended Smart system

Since the initial Smart experiments were in a sense 'unsuccessful', it seemed reasonable to generalize the basic experimental framework in an attempt to determine just what went wrong with the early tests, and to identify indexing search and retrieval methods that would actually prove effective. Accordingly, an extended system was developed with the following capabilities:

- (a) A large number of automatic indexing procedures were made available including operations with automatically generated term associations, and term hierarchies. Furthermore the indexing products could be derived by analysing document titles only, titles and abstracts, or full document texts, and the query-document comparisons could be carried out using a variety of similarity measures¹⁵.
- (b) So-called relevance feedback capabilities were implemented making it possible automatically to generate improved query formulations based on relevance assessments submitted by the users in response to previously retrieved documents. A given user-system interaction could then be carried out in several steps using continually improved query formulations until satisfactory output would be obtained¹⁶.
- (c) Various file organizations could be used including classified, or clustered, collections in which a partial traversal of the stored records would quickly lead to the retrieval of items in areas of interest to the user population¹⁷.

Extensions were also considered by applying the automatic procedures to foreign language documents, and by utilizing bibliographic citations as content identifiers^{18,19}. Eventually, the Smart procedures were compared with the conventional inverted file technologies based on manually assigned keywords to the documents of a collection^{20,21}.

A full discussion of the retrieval results is beyond the scope of this study.

Suffice it to say that a large number of fully automatic retrieval techniques were identified which appeared to be competitive with the more conventional, manual indexing procedures and the inverted file technologies that are conventionally used. Large-scale improvements appeared possible by using the iterative relevance feedback process to reformulate the search requests, and no substantial deterioration results from extending the operations to alien environments such as foreign language materials.

When the automatic procedures incorporated into the Smart system were compared with the manual analysis methodologies used by the Medlars retrieval service at the National Library of Medicine, it was found that the Smart indexing process based on the use of a stored thesaurus produced retrieval results approximately equivalent in terms of recall and precision to those obtainable with Medlars. Using a variety of enhancements such as the automatic relevance feedback procedure, advantages of about 30 per cent could be produced for the automatic Smart system compared to the conventional Medlars process.

Those results turned out to have little immediate impact on operational information retrieval, largely because of the difficulty of rendering believable test results obtained with sample collections of a few hundred documents when the operational environments include several million items. Additional problems are posed by the enormous investments already made in the available commercial systems which make it impossible to contemplate a complete retooling of the kind involved in introducing language analysis methods based on the availability of document abstracts and new file organization methods.

More fundamental complaints were also voiced about the methodologies incorporated into the Smart evaluation system. One of these concerned the necessity to utilize relevance assessments of documents with respect to queries in order to compute recall and precision values. Large-scale studies were made of the relevance assessment process leading to the conclusion that relevance assessments of given documents with respect to particular queries were generally unreliable and not extendible to different system users. Hence it was argued that recall and precision values obtained by averaging the search results over 40 user queries were valid only for the 40 users whose relevance judgements were actually involved^{22, 23}.

Eventually it became necessary to perform a complete study of the question by using a variety of different user populations rendering relevance assessments for the same document collections with respect to the same user queries. It then became clear that the recall-precision results could be expected to remain reasonably invariant with different user populations even though the individual assessments would differ widely²⁴. It was found that substantial agreement existed among groups of assessors for documents retrieved early in a given search that exhibit substantial similarities with the user queries. Those documents are precisely the ones that largely control the shape of the recall-precision curve. There is little agreement for items that are less similar to the queries which therefore appear low down on the output lists; but these documents carry little importance for overall system performance.

Many other objections can be raised about laboratory tests of retrieval systems:

- (a) the use of recall and precision measures to evaluate retrieval systems is objectionable because the user is not interested in merely retrieving relevant items, but rather wants useful items that were previously unknown to him;
- (b) in an iterative feedback environment where search results obtained with earlier query formulations are used to generate improved query statements, the new formulations may retrieve items already seen by the user in an earlier search operation; this circumstance falsifies the evaluation measurements unless special precautions are taken²⁵;
- (c) a number of different strategies may be used to produce evaluation measurements valid for a collection of different users: each user query may be given the same importance regardless of the number of relevant documents the user wishes to retrieve (macroevaluation); on the other hand, each relevant document may be weighted equally, so that a complete response to a query with twenty relevant items would be worth twenty times as much as the response to a query with one relevant item (microevaluation)²⁶.

The list of evaluation problems can be extended, and in principle each objection exhibits merit. In some cases, precautions can be taken to avoid the more obvious pitfalls, and sometimes specific tests can be performed to resolve a particular question, such as the one relating to the variability of the relevance assessments obtained from different user groups. In the case of the Smart environment, many test results are available obtained under differing circumstances with document collections in diverse subject areas and widely differing user populations, and on the whole the results fall into well-defined patterns. By and large, the results do not vary between different document collections, and user groups, and the simpler, better understood methodologies generally prove more effective than more refined procedures that may be difficult to carry out in practice. The methodological objections (other than the obvious ones relating to the restricted collection sizes used in the laboratory) appear to cover second-order effects that are unlikely to invalidate the overall conclusions drawn from the experiments.

15.4 Theoretical insights

The practical effects of the Smart experiments on the operations of most commercial retrieval services may have been relatively small. One can nevertheless point to a number of second-order developments in operational environments: the introduction of global retrieval evaluation measures such as normalized recall and normalized precision²⁶; the adoption of relevance feedback-like procedures in some operational situations²⁷; and the use of automatic document classification²⁸ and automatic term classification methods^{29,30} as an enhancement of the more conventional retrieval methods. The Smart system work has been more influential in creating a new framework for examining the retrieval process. The introduction of the vector processing model, in particular, has led to a re-examination of certain well-established tenets in information and document processing.

Consider, for example, the automatic indexing task. Indexing consists in

the assignment of content identifiers to bibliographic items designed to lead to their retrieval when wanted or their rejection when not wanted. Normally, the indexer considers each item in isolation and assigns content terms that are related in some sense to the document content. This procedure may not lead to effective retrieval, because the choice of appropriate index terms depends not only on the contents of each individual document, but also on the contents of all other documents in the collection. For example, the term 'computer' may be appropriate in identifying a document entitled 'Uses of Computers in Medicine' if such an item is placed in a collection of medical items, most of which will necessarily be unrelated to computers. 'Computer' would be a poor choice for that same document if the item were to be placed in a computer science collection, because then all other documents are also computer-oriented.

Thus, indexing implies the assignment of content identifiers to documents that are capable of reflecting the document content in some sense, *and* that distinguish the items from each other. In the vector space environment, distinguishing the items implies decreasing their similarities, or increasing their mutual distance in the space.

The requirement to create a document space that is spread-out, that is, where the distances between document vectors are as large as possible, leads to the assignment of term importance values, or term weights to the content identifiers used for indexing purposes. One such indication of term importance is the term **discrimination value** which measures the ability of a term to spread out the document space when assigned to the documents of a collection³¹⁻³³. In the absence of information about the actual term relevance, one can relate the term discrimination value to various occurrence frequency characteristics of the terms in a collection^{34,35}. It turns out that the best terms will be medium-frequency terms that are not assigned to too many documents in a collection nor to too few because high-frequency terms assigned to many items in a collection render the document vectors more similar to each other, thereby compressing the space, and rendering it difficult to retrieve the individual items when wanted; low-frequency terms, on the other hand, are assigned to so few documents that their overall effect is not sufficiently felt. When medium-frequency terms are used, those items to which they are assigned are rendered more similar to each other, but at the same time the differences between such items and the remainder of the collection will be increased. This is symbolically illustrated in the document space representation of *Figure 15.1*, where each x denotes a document, and the distance between two x 's is assumed to be inversely related to the similarity in the respective document vectors.

The space alteration of *Figure 15.1* is obviously desirable under the assumption that the items to which term k is assigned will prove jointly relevant to the users' information requests: these items are made similar to each other rendering them easily retrievable together and thus producing high recall; at the same time, these items are distinguished from the remainder of the collection, which leads to high precision and to the correct rejection of the extraneous items.

The term discrimination model is used to generate an automatic indexing system in which the discriminating medium-frequency terms serve directly for indexing purposes. The high-frequency terms that compress the document

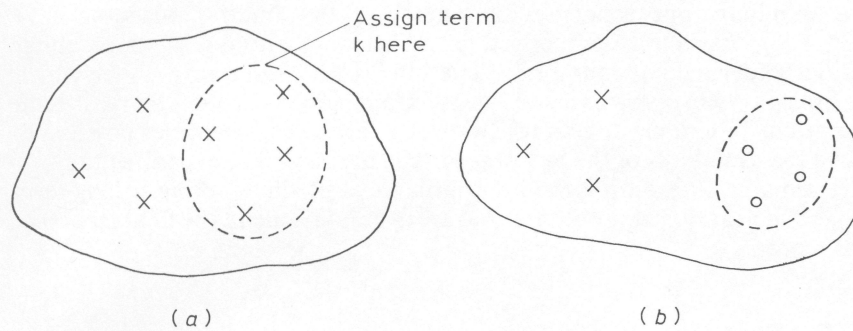


Figure 15.1. Basic document space alteration. (a) Before assignment of term k ; (b) after assignment of term k

space when used for indexing purposes must be rendered more specific: their frequency of assignment can be decreased by incorporating the terms into **term phrases**, and assigning the phrases as content identifiers (for example, instead of using 'computer' as an index term, one could form the phrases 'computer programmer', or 'computer hardware', or 'computer security'). Low-frequency terms, on the other hand, can be broadened by incorporating the terms into **thesaurus classes** consisting of groups of related or synonymous terms. Each thesaurus class necessarily exhibits a higher assignment frequency in a collection, than the individual terms included in a thesaurus class³⁶.

The vector space model of information representation and retrieval is thus capable of assigning a specific interpretation to well-known intellectual content analysis aids such as term grouping and thesauri, and this role is different from the standard semantic functions of such devices in linguistics. When relevance information of documents with respect to search requests is available in retrieval (as is the case in many systems that provide user-system interaction), then a term relevance factor known as **term precision** can be computed as the proportion of relevant items containing a given term to total number of items (or to number of non-relevant items) containing the term. It is clear that terms with a high precision factor are capable of distinguishing the relevant items from the non-relevant ones; the neutral term discrimination weights can then be replaced by term precision weights. It has been shown that a term weighting system based on the use of term precision is theoretically optimum for the binary-independent retrieval model (where binary weighted terms are independently assigned to queries and documents)^{37,38}. Furthermore, a good deal of experimental evidence is available demonstrating the usefulness of the term relevance factors even in cases where the binary independent model does not strictly apply^{39,40}.

For the binary-independent model which is relatively easy to treat mathematically, various Smart procedures can also be shown to be formally effective under specified circumstances. Thus an effective cluster search method is available which is capable of concentrating the search effort in the most productive areas of a classified collection⁴¹. Formally effective document and query vector alteration methods have also been studied, including in particular the Smart relevance feedback process^{42,43}.

In summary one concludes that many of the Smart procedures have interesting theoretical properties, in addition to proving effective under various experimental conditions. The intellectual framework under which the Smart system operates makes it easy to add new procedures and to extend operations in various directions. In quite a few cases it becomes possible to prove the usefulness of the techniques formally as well as experimentally.

It remains to examine the appropriateness of undertaking a long-term project such as Smart in the retrieval area. This is done in the final section of this report.

15.5 Concluding remarks

It is hardly necessary to point out that the Smart system design carries with it great advantages if one aims at constructing a flexible environment for retrieval system experimentation. Whereas in normal environments, it becomes necessary to retool to begin each individual experiment, the Smart system has made it possible to carry out hundreds of different experiments without substantial overhead or expense in program modification or collection preparation. Such a flexible environment is to some extent bigger than the sum of its parts: after using the system for a while one sees things fall into place; often one can anticipate the evaluation results before actually seeing them, and one obtains an intuitive feeling for the operations of a retrieval system. It is then possible to obtain substantial returns from a continuing experimental project, in return for the substantial investment that is necessary in building and maintaining the system over many years.

Normally, an experimental system is considered useful because the experimental results can help confirm a variety of formal theories and abstract models for a given process or system of procedures. The Smart system experiments have in fact been initiated in an attempt to confirm a variety of theories about the content analysis problem. When an experimental system is sufficiently flexible it may also be useful 'in reverse'. That is, the test results can help in formulating theories, and formal proofs can sometimes be generated to describe precisely the conditions under which a given experimental process is expected to be useful. Formal results obtained after the fact have thus helped in rendering the Smart test results plausible in areas such as term frequency weighting, term precision weighting, document clustering, and relevance feedback.

In addition the Smart system results have led at least to a rethinking about, and sometimes to actual modifications of existing retrieval procedures. Since so many different methodologies were actually subjected to intensive tests in areas such as document input, indexing, classification, document-query comparison, output ranking and display, query reformulation, and so on, the Smart system has something to say in most areas relating to information system design. As a result selected methods that are easy to implement and apparently most productive (term weighting, relevance feedback, etc.) have in fact found their way into a number of operating environments.

What about the drawbacks of a large and continuing experimental project? Obviously one must be careful about the initial design and about the claims one makes about the results. It is easy to go off on a tangent and to get stuck

in a morass of one's own creation. If the system is misdesigned and does not adequately reflect any part of the real world, the evaluation results themselves will likely prove to be useless. This point of view has been espoused most cogently by L. B. Doyle in an early book review⁴⁴:

'A comment is needed about the Smart system . . . The word "system" is misleading. It is really a chemistry laboratory for retrieval principles and procedures . . . it is a *tour de force* in experimentation in the documentation area, the like of which is seldom seen . . . My only reservation about Smart is that it may not be doing the right kind of chemistry—but then hardly anyone is . . . The aspect adjudged most negative is that so much research should have been done by one party under a suboptimized set of assumptions. . . .'

It is now unfortunately too late to ask the author of the foregoing quote to explain these statements—the only comment actually made by Doyle raises the question of 'what good is a retrieval system when nine-tenths of the possible users use the telephone instead?'—a statement that is surely less appropriate in 1979 than it was when written in 1969. But obviously the reviewer's principal contention is certainly correct: if the assumptions in an evaluation system are suboptimized, the results may not be worth a great deal.

How then do the Smart assumptions relate to reality? In principle, many questions can be raised about the appropriateness of the basic model, quite apart from the problem specifically due to the restricted experimental environment. Thus the vector space model may be questioned based on the fact that the scope (as opposed to the subject area) of an item cannot be represented by a simple vector length and direction. In particular, two items might cover the same subjects and hence be represented by identical vectors, yet the topic areas could be treated narrowly in one case and broadly in the other. The suggestion is therefore made that bibliographic items should be represented by vectors supplemented by scope or extension measures, instead of by vectors alone as in Smart⁴⁵.

Other problems, already mentioned in part, concern the implicit assumptions of term independence in the vector processing model, that is, the partially false notion that content identifiers occur independently of each other in document and query vectors⁴⁶. Independence among evaluation parameters is also assumed by certain statistical tests used in Smart to assess the significance of the evaluation results. Additional methodological objections are easy to find in a computer environment comprising many tens of thousands of processing steps.

Some of these comments are formally correct; a laboratory model by its very nature can never fully reflect the real life conditions. The question is whether the deviations are sufficiently serious to affect the usefulness of the model. So far as Smart is concerned, the first-order characteristics of the real world are believed to be properly represented—the subject content of a document represented by a vector in multi-dimensional space is more important than its extension, and to a first-order approximation, the terms used to characterize the documents are indeed independent. Furthermore, while the formal proofs of effectiveness of some Smart procedures are applicable only in restricted environments (binary vectors, term independ-

ence, inner product similarity function, etc.), one may expect that the system operations are representative of a much wider area outside the formal limits.

After more than ten years of experimentation with Smart, the main problem is not apparently the fact that Smart may have been 'doing the wrong kind of chemistry', but rather the fact that laboratory-type chemistry is not quite the same as production work in a chemical factory; which does not imply of course that laboratory work is useless or unnecessary.

References

1. LUHN, H. P. The automatic derivation of information retrieval encodements for machine-readable texts. In: *Information Retrieval and Machine Translation*, (Ed. A. Kent), Part 2, Interscience, New York (1961)
2. LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information, *IBM Journal of Research and Development* **1**, 309-317 (1957)
3. SWANSON, D. R. Searching natural language text by computer, *Science, New York* **132**, 1099-1104 (1960)
4. MONTGOMERY, C. and SWANSON, D. R. Machine-like indexing by people, *American Documentation* **13**, 359-366 (1962)
5. SALTON, G. Manipulation of trees in information retrieval, *Communications of the ACM* **5**, 103-114 (1962)
6. SALTON, G. *Dynamic Information and Library Processing*, Chapter 6, Prentice-Hall, Englewood Cliffs, N.J. (1975)
7. ATHERTON, P. *A Proposed Standard Description for Reporting Evaluation Tests of Retrieval Systems*, presented at Seventh Institute on Information Storage and Retrieval, American University, Washington, D.C. (1965)
8. SPARCK JONES, K. and BATES, R. G. *Report on a Design Study for the Ideal Information Retrieval Test Collection*, British Library Research and Development Report 5428, Computer Laboratory, University of Cambridge (1977)
9. FAIRTHORNE, R. A. Basic parameters of retrieval tests, *Proceedings of the American Documentation Institute* **1**, 343-346, American Documentation Institute, Washington, D.C. (1964)
10. SALTON, G. The evaluation of automatic retrieval procedures—selected test results using the Smart system, *American Documentation* **16**, 209-222 (1965)
11. SALTON, G. and LESK, M. E. The Smart automatic document retrieval system—an illustration, *Communications of the ACM* **8**, 391-398 (1965)
12. SALTON, G. Designing automatic information systems: results obtained with the Smart programs, *Social Science Information* **6**, 111-117 (1967)
13. CLEVERDON, C. W. and KEEN, E. M. *Factors Determining the Performance of Indexing Systems*, Vol. 2—*Test Results*, Aslib Cranfield Research Project, College of Aeronautics, Cranfield, England (1966)
14. BOURRELLY, L. and CHOURAQUI, E. *Le Système Documentaire Satin 1, Vol. 1, Description Générale et Manuel d'Utilisation*, Centre National de la Recherche Scientifique, Paris (1974)
15. SALTON, G. and LESK, M. E. Computer evaluation of indexing and text processing, *Journal of the ACM* **15**, 8-36 (1968)
16. IDE, E. and SALTON, G. User controlled file organization and search strategies, *Proceedings of the ASIS National Conference*, Vol. 6, pp. 183-191, Greenwood Publishing Corp., Westport, Connecticut (1969)
17. SALTON, G. *Search Strategy and the Optimization of Retrieval Effectiveness*, *Mechanized Information Storage, Retrieval and Dissemination*, (Ed. K. Samuelson), pp. 73-107, North-Holland, Amsterdam (1968)
18. SALTON, G. Experiments in multilingual information retrieval, *Information Processing Letters* **2**, 6-11 (1973)
19. SALTON, G. Associative document retrieval techniques using bibliographic information, *Journal of the ACM* **10**, 440-457 (1963)
20. SALTON, G. Recent studies in automatic text analysis and document retrieval, *Journal of the ACM* **20**, 258-278 (1973)

21. SALTON, G. A new comparison between conventional indexing (Medlars) and automatic text processing (Smart), *Journal of the American Society for Information Science* **23**, 75–84 (1972)
22. CUADRA, C. A. and KATTER, R. V. Opening the black box of relevance, *Journal of Documentation* **23**, No. 4, 251–303 (1967)
23. REES, A. M. and SCHULTZ, D. G. *A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching*, 2 Vols, Center for Documentation and Communication Research, Case Western Reserve University (1967)
24. LESK, M. E. and SALTON, G. Relevance assessments and retrieval system evaluation, *Information Storage and Retrieval* **4**, 343–359 (1968)
25. SALTON, G. Evaluation problems in interactive information retrieval, *Information Storage and Retrieval* **6**, 29–44 (1970)
26. SALTON, G. The evaluation of computer-based information retrieval systems, *Proceedings of the FID Congress*, pp. 125–133, International Federation for Documentation (1965)
27. VERNIMB, C. Automatic query adjustment in document retrieval, *Information Processing and Management* **13**, 339–353 (1977)
28. CROFT, W. B. Clustering large files of documents using the single link method, *Journal of the American Society for Information Science* **28**, 341–344 (1977)
29. ATTAR, L. and FRAENKEL, A. S. Local feedback in full text retrieval systems, *Journal of the ACM* **24**, 397–417 (1977)
30. DOSZKOC, T. E. AID—an associate interactive dictionary for on-line searching, *On-Line Review* **2**, 163–173 (1978)
31. SALTON, G., YANG, C. S. and YU, C. T. A theory of term importance in automatic text analysis, *Journal of the American Society for Information Science* **26**, 33–44 (1975)
32. SALTON, G., WONG, A. and YANG, C. S. A vector space model for automatic indexing, *Communications of the ACM* **18**, 613–620 (1975)
33. SALTON, G., YANG, C. S. and YU, C. T. Contributions to the theory of indexing, *Information Processing 74, Proceedings of IFIP Congress* (Ed. J. L. Rosenfeld), pp. 584–590, North-Holland, Amsterdam (1974)
34. SPARCK JONES, K. A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation* **28**, 11–21 (1972)
35. SALTON, G. and YANG, C. S. On the specification of term values in automatic indexing, *Journal of Documentation* **29**, 351–372 (1973)
36. SALTON, G. and WONG, A. On the role of words and phrases in automatic text analysis, *Computers and the Humanities* **10**, 69–87 (1976)
37. ROBERTSON, S. E. and SPARCK JONES, K. Relevance weighting of search terms, *Journal of the American Society for Information Science* **27**, 129–146 (1976)
38. YU, C. T. and SALTON, G. Precision weighting—an effective automatic indexing method, *Journal of the ACM* **23**, 76–88 (1976)
39. SALTON, G., WONG, A. and YU, C. T. Automatic indexing using term discrimination and term precision measurements, *Information Processing and Management* **12**, 43–51 (1976)
40. SALTON, G. and WALDSTEIN, R. K. Term relevance weights in on-line information retrieval, *Information Processing and Management* **14**, 29–35 (1978)
41. SALTON, G. and WONG, A. Generation and search of clustered files, *ACM Transactions on Database Systems* **3**, 321–346 (1978)
42. YU, C. T., SALTON, G. and SIU, M. K. Effective automatic indexing using term addition and deletion, *Journal of the ACM* **25**, 210–225 (1978)
43. WU, H. *On Term Distribution, Space Density and System Performance*, Department of Computer Science, Cornell University (1979)
44. DOYLE, L. B. Book review of automatic information organization and retrieval by G. Salton, *Computing Reviews* **10**, 271–272 (1969)
45. DE Solla Price, DEREK, private communication
46. VAN RIJSBERGEN, C. J. A theoretical basis for the use of cooccurrence data in information retrieval, *Journal of Documentation* **33**, 106–119 (1977)