A NEW TYPE OF INFORMATION RETRIEVAL SYSTEM*

H. Joel Jeffrey
Systems & Information Science, Vanderbilt University
Nashville, Tennessee 37235

ABSTRACT

In the period 1964-1968, Peter G. Ossorio [10,11,12] developed and tested, on a pilot study basis, a new approach to the problem of automatic document retrieval. Ossorio's studies were entirely successful, as pilot studies, and show the feasibility of using his approach to produce a new kind of retrieval system.

These retrieval systems do not operate by word matching. The basic approach is to simulate the judgement of competent human judges of the conceptual content of each document, and the request. This judgement is then used to retrieve those documents with conceptual content most similar to that of the request.

Each document is processed only at the time it is added to the data base, in time linear in the number of words in the document that the system recognizes. The retrieval request is in ordinary English. Time for retrieval is linear in the number of documents on file. Documents are retrieved in order of similarity of conceptual content to that of the request. The system works, in certain respects, better on full text documents, providing better descriptions of document content, and more detailed cross-indexing.

The new type of system shows a number of interesting features. Among these are:

(1) Much better performance than systems using the old techniques;

(2) Faithful representation of the judgement of the person(s) whose judgement is being simulated, thus providing the possibility of individualized retrieval systems;

(3) Ability to explain to a user why it retrieved certain documents, and not others. With this information, the user can alter his request, or instruct the system to judge things differently;

(4) Automatic recognition of requests the system cannot properly handle;

(5) Sub-documentary indexing reflecting heterogeneity of material. As is often the case with a new paradigm, Ossorio's work raises at least as many questions as it answers. This paper presents the new approach, and the results of some first explorations in the new field.

INTRODUCTION

When a person retrieves information in response to a request, he uses his judgement to retrieve the information which is closest to what he takes to be the desired conceptual content. He may use various criteria in making his judgement. Subject matter relevance is one such criterion. Another is type of content, e.g., mathematical vs. physical. A paradigm case is a library staffed by a competent, knowledgeable librarian.

*This work was supported by National Science Foundation Grant Eng75-10492.

The ability to make content judgements allows the librarian to aid the user in a number of ways:

(1) He can retrieve documents in order of closeness of content;

(2) He can retrieve a part of some document, e.g., a paragraph from an article, which is relevant to the request, although the larger document is not;

(3) He can negotiate with the user, emphasizing or de-emphasizing some area of content at the user's direction.

An automatic retrieval system operates in a far different manner. To date, automatic systems have lacked the ability to make the judgements necessary to retrieve as a person does. In lieu of judgement, information retrieval has been based on the fundamental approach, i.e., paradigm [1], of word matching.

The fundamental difficulty with the keyword paradigm is that the acceptability of a document as a response to a request depends, in almost all cases, on the content of the document, not on the words in it. When a person uses a library, retrieved documents may in fact not have a single word in common with the words of the request. When one is limited to word processing, this leads to the enormous problem of knowing how to match a document and a request containing different terms.

As might be expected from this fact, current retrieval systems have a number of difficulties:

(1) Poor performance. Generalizations of performance figures are difficult and somewhat crude. However, as a rule-of-thumb estimate, typically % recall + % precision = 100%. Salton, for example, cites the following as typical figures: 10+95, 30+80, 40+70, 60+35, 80+20. [2] Similar results are very common in the literature. [3,4,5,6] These results are, simply, not very good. 50% recall, 50% precision is clearly unsatisfactory for many purposes.

(2) Operation on full text is very difficult and slow, being hampered by time to process documents for indexing;

(3) The system must either restrict the vocabulary available to the user, or else deal with severe thesaurus problems and slow response;

(4) These restrictions force users to acquire considerable sophistication in the operation of a particular system, or else employ a specially trained operator. This is the case, for example, in an operational system for the retrieval of medical literature, MEDLARS. [7]

THE NEW TYPE OF SYSTEM

It has always been assumed that retrieval by analysis of conceptual content must await solution of the problem of natural language understanding by computer, and there are strong reasons for pessimism as to a solution for that problem. [8,9]

That assumption is false.

We can, in actual fact, do content judgements automatically without having to solve the language understanding problem. In the period 1964-1968, Peter G. Ossorio developed, and demonstrated the practical utility of, a method by which we can bypass the problem of natural language understanding, to produce retrieval systems that can make and use conceptual content judgements. [10,11,12]

Systems based on this approach retrieve as a person does. Specifically, when the system receives a request, it makes a judgement of its conceptual content, and retrieves those documents whose content most closely matches that of the request. The problem of matching words used in a document and the request never arises.

We call this approach the judgement paradigm.

Judgement-based systems have several highly desirable characteristics:

(1) The retrieval request is in ordinary, unformatted English, with no vocabulary restrictions.

(2) The system delivers documents to the user in order of relevance.

(3) Performance is much better than that offered by the traditional type of system. We are currently achieving recall plus precision of 160-170%, and there is very strong evidence that this will soon be improved, with little effort, to 190% or higher.

(4) System operation is extremely fast, with (a) time to add a document to the file linear in the number of words in the document, and (b) time to retrieve documents linear (with a small constant) in the number of documents on file.

(5) Operation on full text is no more difficult than on abstracts, and system performance is, in fact, enhanced with full text, with more detailed cross-indexing and enhanced recall and precision.

(6) The system is rational, i.e., it can report to the user its content judgement, and alter it as instructed. This is far different, and vastly more powerful, than interactive relevance feedback as it is currently done. [2,13,14]

(7) Systems can be constructed to reflect the information needs and desires of a single individual. Such a "personalized" system will often be much superior for a user.

CONSTRUCTION OF A JUDGEMENT-BASED SYSTEM

Construction of a retrieval system with judgement capability proceeds as follows:

(1) Select the area--the subject matter fields, or topics, to be covered. All questions of overlap of fields, or any of the myriad relationships that might hold between fields, are ignored.

(2) Select the System Vocabulary--words and phrases from the area to be covered.

(3) Obtain, from expert human judges, ratings of the relevance of each item of the System Vocabulary with respect to each subject matter field. Judges rate the terms as follows:

(a) Irrelevant.
(b) Possibly relevant. The item might have some relevance to the field.
(c) Peripheral. The item has some relevance, but is basically peripheral to this field.
(d) Relevant. The item is definitely part of the field.
(e) Highly significant. The item is an important concept in the field.

Within each of the categories b-e, the judge specifies more, or less relevance. The expert judgements are expressed as numbers on a 0-8 scale. Zero indicates irrelevance, 8 highly significant relevance, etc.

For simplicity, we have phrased the above description in terms of subject matter relevance. The technique is not limited to this type of judgement. Any type of content judgement may be used. If, r example, one wants the system to judge properties, then instead of subject matter fields, we begin with descriptions such as "x is mathematical", or "x is physical". Judges then rate the degree to which each description applies to each item.

At this point, all human intervention is finished. No further judgements are used, and all processing is completely automatic.

(4) We now form a matrix, with one column per topic, and one row per term. Due to overlapping topics, this matrix contains a great deal of redundancy. If we do not deal with this redundancy, it will cause problems when we try to use the information.

Formally, if we have n topics and t terms, then we have t n-vectors. However, the overlap of topics results in a dimensionality of less than n, in general. We solve the problem of the redundant information by finding an orthogonal basis for the vector space. We do this with a technique common in the social sciences:

(5) We intercorrelate and factor-analyze the matrix. Each common factor then represents a type of conceptual content, distinct (due to the orthogonality) from all other factors.

The basis is now all measurable common factors plus a unique factor for any topic not well-represented in the common factors. Such topics are those with a substantial portion of unique content.

We call this vector space a judgement space.

(6) We now calculate the location of each term of the System Vocabulary in the space, i.e., its judgement vector.

(7) Calculate a judgement vector for each document of the data base, by a function of the judgement vectors of each vocabulary item found in the document. A variety of functions are possible, of course. We are currently using:

$$D_k = \frac{A + B}{2}, \text{ where}$$

$$A = \frac{\sum_{i=1}^{n} t_{ik}}{n}$$

$$B = \frac{\prod_{i=1}^{n} t_{ik}}{\sum_{j=1}^{r} \prod_{i=1}^{n} t_{ij}}$$

where: k = k-th component of the judgement vector
n = number of terms in the document
r = number of axes

Orthogonality has important consequences here: to calculate a document's position on axis k, only the k-th component of each term vector is used.

At this point, the data base processing is complete. Documents are not reprocessed for a request.

(8) To retrieve a document, treat the request as any other document, and produce a judgement of it. Then, retrieve documents in order of closeness in the vector space.

The key point here is that closeness in the space reflects closeness in conceptual content.

We can now see the source of the rationality mentioned earlier: the system can report to the user its judgement of request content, and then alter that judgement at the user's direction.

PRACTICALITY

There are a number of issues pertaining to the practicality of this technique. We do not have space for more than a very brief discussion here. Most of the basic questions were settled by Ossorio. His work and ours shows that the technique is completely practical and works well.

(1) Judgements are reliable, although they will vary from person to person, depending on point of view. As a result of this fact, we now have the capability of constructing systems that reflect a single individual's point of view, as well as a more normal "consensus" system. Such systems would appear to be quite valuable in certain uses.

(2) The factors remain stable across different selections of vocabulary. There are no problems whatsoever with the factor analysis itself.

(3) The amount of human effort, in giving term ratings, is entirely practical (although certainly not trivial).

(4) Time necessary for document processing and retrieval is entirely practical. In this regard, we note that, due to the original numerical ratings, differences in distance of less than 0.1 (as a very conservative estimate) are not significant. As a result, sorting of documents by distance is linear.

RESULTS

In our opinion, we have gone as far as we can go with thought experiments and pilot studies. What is needed now is experience with judgement-based systems--experience by real users with need of a good retrieval system.

We are currently building such systems. A computer science system is currently operational. The following results are fairly typical of performance at this time:

Request 1: "I am interested in the decidability of the equivalence problem for deterministic pushdown machines.:
Request 2: "What do you have on paging?"
Request 3: "What do you have on graph theory?"
Request 4: "What do you have on operating systems?"
Request 5: "What have you got about programming languages?"

	Retrieval cutoff at distance of 4.6.		Retrieval cutoff after 3 consecutive irrelevant documents (last 3 not included in figures).	
Request	Precision	Recall	Precision	Recall
1	1/1	1/1	1/1	1/1
2	8/16	8/9	6/10	6/8
3	15/18	15/16	16/19	16/16
4	16/16	16/36	36/37	36/36
5	18/34	18/18	14/17	14/18
	Average Precision:	77%	Average Precision:	85%
	Average Recall:	86%	Average Recall:	90%

The system so far is quite crude, with a small vocabulary, no heirarchical subspace structure, and no negotiation capability. Each of these, particularly vocabulary enlargement, will greatly improve performance. (Here we see a direct effect of content retrieval--adding vocabulary is entirely uncomplicated, and never degrades performance.)

Thus, while this performance compares very favorably with that of systems based on the old paradigm, we view this as a lower bound on what we expect to have very soon.

FUTURE WORK

We have referred to judgement-based systems as a new paradigm for information retrieval. It is the mark of a new paradigm in a field [1] that, with a new paradigm, work in the field is not simply devoted to finding new answers to old questions. Rather, the research questions themselves change. That appears to be the situation here. Most, if not all, of the issues we are exploring simply did not exist before. Further, virtually all of the work in automatic indexing and document clustering is at best peripherally relevant to our work. Systems with judgement capability differ radically and totally from those without it.

In addition to purely technical questions, such as which document locating functions, and which distance functions, are better, we are currently working on individualized retrieval systems, hierarchically structured subspaces, techniques for using content judgement as a context to alter the processing of marginal terms, and retrieval of non-linguistic information. In each case, our goal is the same: to gain real-world experience with an aspect of a new technology for information retrieval.

REFERENCES

1. T.S. Kuhn, The Structure of Scientific Revolutions, University of Chicago Press, Chicago, Illinois, 1975.

2. G. Salton, Dynamic Information and Library Processing, Prentice-Hall, Englewood Cliffs, N.J., 1975.

3. F.W. Lancaster and E.G. Fayen, Information Retrieval On-Line, Melville Publishing Co., Los Angeles, California, 1973.

4. F.W. Lancaster, Evaluation of the MEDLARS Demand Search Service, National Library of Medicine, Bethesda, Maryland, 1968.

5. F.W. Lancaster, "Evalutation of On-Line Searching in MEDLARS (AIM-TWX) by Biomedical Practitioners", in Lancaster [3] above.

6. F.W. Lancaster, R.L. Rapport, and J.K. Perry, "Evaluating the Effectiveness of an On-Line, Natural Language Retrieval System", Information Storage and Retrieval, Vol.8, No.5, pp.223-245.

7. The Principles of MEDLARS, National Library of Medicine, Bethesda, Maryland, 1970; available from Superintendent of Documents, Washington, D.C.

8. G. Salton, "Automatic Text Analysis", Science, Vol.168, No.17, April 1970, pp.335-343.

9. T. Winograd, Procedures as a Representation of Data in a Computer Program for Understanding Natural Language, revised version of Ph.D. Dissertation, Department of Mathematics, August 24, 1970, M.I.T., Cambridge, Massachusetts.

10. P.G. Ossorio, Attribute Space Development & Evaluation (RADC-TR-67-640), Rome Air Development Center, Rome, N.Y., 1968.

11. P.G. Ossorio, Dissemination Research (RADC-TR-65-314), Rome Air Development Center, Rome N.Y., 1965.

12. P.G. Ossorio, Classification Space Analysis (RADC-TR-64-287), Rome Air Development Center, Rome N.Y., 1964.

13. G. Salton, "Recent Studies in Automatic Text Analysis", Journal of the ACM, Vol. 20, No.2, April 1973, pp.258-278.

14. G. Salton, "Dynamic Document Processing", Communications of the ACM, Vol.15, No.7, July 1972, pp.658-668.