THE BASIS FOR A THEORY OF INFORMATION RETRIEVAL

J. Farradane
School of Library & Information Science,
University of Western Ontario, London, Ontario, Canada N6A 5B9.

Any theory which attempts to explain and predict the behaviour of a system, or to form the basis for construction of a system, must be founded on *evidence* of the characteristics and behaviour of that system, or of its components. Information retrieval systems have been so much of an empirical nature that there has been little evidence on which to base reliable theory. Some systems have been constructed on principles of linguistics, such as De Saussure's theory in Gardin's "Syntol" system, or on the assumption of the interrelation of words by their frequency of occurrence or co-occurrence, but both such approaches assume that linguistic structure always has a clear relation to meaning. Meaning originates in thought, but language is only a surrogate, and often a poorly used surrogate, for such thought.

Many theoretical studies have been based on probability theory applied to retrieval possibilities, but such theories have had little basis of experimental facts, and have not been developed to make predictions for which experimental verification has been sought. There are in fact several stages of information retrieval where one might expect a special bias to operate.

It is possible to make a good *fact*-retrieval system from records of invariant data such as, say, physical properties, with a limited number of standard terms, as in a data handbook, but this is not the type of system in general use. Most retrieval systems are intended to cover much more diffuse types of information, and have been based on the use of keywords and/or descriptors, and their connection by elementary Boolean algebra in the putting of questions. Neither of these correctly or adequately expresses meaning between terms, especially if the terms have been inadequately standardized or selected. There have consequently been a number of empirical attempts to overcome false drops and other errors by such means as the use of 'or'-terms, truncated terms, roles, links, word co-occurrence rules, weighting, feedback, etc., but without any real success or consistency.

It is therefore important, for the development of valid theory, to determine the facts involved, that is, the basic characteristics of 'information', and of the way it is produced, recorded, organized in systems, and received by users. This is not an easy task. It starts with examining thought and its expression in the surrogate of language or other media, whereby it becomes 'information'.[1] I define information as such a surrogate of thought used for communication. Indexing, and other manipulations of the information, also introduce changes in the representation of thought, and the retrieval system has to deal with the indexed form, and allow for these variations. Full-text searching appears to avoid the indexing, though with much redundancy, but is still carried out by simple coordinate searching, which does not represent meaning with any real accuracy; furthermore, the author's language may be far from standard. Searching methods have been only simple matching processes, but I will show that several other routines can usefully be introduced.

Next, we need to elucidate what is involved in the user's decisions of 'relevance', i.e. the interaction of information from retrieved material with the user's prior knowledge, or with the intention of his question, which can change with each acquisition of new data. This brings us back full circle to the nature of thought.

Finally, there is the problem of measuring the working of the system. The usual measures now applied (recall, precision, etc.) are considered to be related to the performance of the system as a whole, but are in fact based only on the user's estimates of relevance. The user's judgment is however a little understood factor, though nevertheless the controlling factor, and this does not allow us to examine separately the efficiencies of other parts of the total system.

Two procedures involve fallacies which have not been sufficiently appreciated:

1. Boolean algebra is a mathematical representation of logical interrelations between *statements* or *classes*, but does *not* correctly apply to relations between single words or concepts. It can be pictured as applying to some computer operations, and this has made it appear suitable for handling questions. In fact, all it achieves is the finding of documents whose indexing includes

two or more words simultaneously; the use of 'or'-logic allows for search of alternative words, but this clearly alters the question. Further, the Boolean 'not' is well-known to be much more misleading than it appears to be. The impression that Boolean connectors provide meaning is therefore erroneous; the use of unspecific 'links' or arbitrary 'roles', is equally misleading.

2. Another procedure to be questioned is the use of 'recall' and 'precision' as measures of efficiency, and, in particular, the much quoted assumption that they are inversely related. Of course, both ratios involve user judgments of relevance, which are subjective and not very consistent, and depend on the level of knowledge of the user. Judgments by others than the original questioner may also vary considerably, as has been shown by tests even among the co-workers of a single research group. As individual (unrelated) measures of practical output, the ratios may be helpful to a given user, but they cannot be used as measures of the total system. Furthermore, the claimed inverse relationship between recall and precision has in fact been an artifact of the method of testing. In Cleverdon's work, the inverse relationship was artificially introduced by plotting the averaged answers to different 'levels of coordination' of the questions, i.e. with fewer terms than the total question; this was done in order to obtain a curve, and not just single points of recall-precision pairs. But the answers to the questions with fewer terms were however judged by the answers to the total questions; the questions with fewer terms are however wider in scope, and are thus different, and so answerable correctly by other documents, which are however not accepted. Precision is therefore bound to decrease at lower levels of coordination; recall may increase by the finding of documents which were not initially indexed in sufficient detail. Such curves are therefore meaningless as measures of the system, and are only a demonstration of the effect of altering the question in relation to a fixed set of answers!

Salton's method of ranking the output documents, and determining the recall and precision at each occurrence of a relevant document in the list, similarly forces an inverse relationship, since if any non-relevant documents appear in the output, the precision *must* decrease as the recall increases.

The true relationship between recall and precision, if in fact there is *any* relationship, is as yet unknown. In my experiments it has been found that both high recall and high precision can be obtained together. Furthermore, as Fairthorne, and also Swets, have pointed out, these measures, taken alone, leave out the important figure of the file size in the system. Fallout, which they advocate instead of precision, does incorporate the file size, but is too insensitive a figure unless the file size is small, or the system is very inefficient.

The process of indexing also needs investigation. The indexer's level of knowledge, and his ability to represent information by suitable words, are involved. A thesaurus is the usual aid here, but most thesauri are incomplete, and of course soon get out of date; they include far too many arbitrarily compound terms, which are unnecessary if the correct interconnections between individual simpler terms could be recorded. An author has converted his thought to language; the indexer has to understand this, that is to reconvert it into *his* thought in order to represent it again by indexing terms. The indexing process thus requires special controls.

It seems desirable to examine each stage of the information retrieval process separately, if possible, and then to investigate the interactions between these stages, before one can evaluate a whole system or compare it with others, so that improvements can be made.

Such problems have been under investigation by me for several years. Starting from a basis of classification theory, it was found that there was adequate evidence from the psychology of thinking to identify a system of nine categories of relation between terms (concepts), sufficient to enable complex subjects, in any field, to be represented by linear or two-dimensional structures of concepts with interposed (coded) relations; these structures are very exact in meaning and (with some training) reproducible, and have self-regulating characteristics which guide the indexer. This method has been called Relational Indexing, and has been fully described elsewhere.[2] It must be combined with standardization of terminology, and its presentation according to several rules; thus all verbs (processes, actions) are used in the gerund form, ending in 'ing', so that, for example, 'government' is not confused with 'governing'. Different tenses are expressed by diferent relations. Compound concepts are avoided as much as possible; all concepts must basically be nouns or verbal nouns; most compound ideas can be constructed by suitable relational combinations, but if adjectives are needed they can be added *after* a noun *and* a comma, and the program can then search for the noun with or without the adjective, e.g. 'ultraviolet light' is indexed as 'light, ultraviolet', which can be found in a search for 'light' alone if so required. Of course, searches also identify the interposed relations. In making a data base, a word list (with frequencies of use) can be obtained from the computer at intervals, to enable a check to be made on undesirable synonyms and other errors.

A demonstration that the information has not been distorted by the relational representation has been achieved by a computer program which reconverts the diagrams into readable statements in English, including prepositions, where necessary, in place of the relational symbols, and these statements are equivalent to the original information; these statements, with controlled permutations and suitable indentations, can then be arranged alphabetically to provide a reasonable printed subject index.[3]

On investigation of users, it was found that the user often unwittingly makes condensations in his formulation of a question, perhaps omitting some implied terms, and rules were found for imitating such variations, so that the initial more detailed indexing can be reduced by rule to the possible condensed versions of a question.[4] The computer can produce these 'logical jumps' automatically on request, and *adds* the results to the indexed formulation, without replacing it.

These logical jumps have been fully validated in later work.

Tests have been made with three different specially prepared data bases, and these have given very good results[5]; precision has been very high (as much as 93%) and recall has been at least 70%. The lower recall values seem to be due mainly to user uncertainty. The measures of recall and precision have been used until better measures can be determined. The user's questions are of course indexed in the same way as are the documents, and a match can be made of the question with any part of the document indexing. It is desirable and useful for questions to be put with specific terms, although general terms can be allowed for in the indexing.

New diagnostic programs have been developed to satisfy new research requirements. All procedures have been incorporated, and the program can also yield a printout of the progress of each procedure, so that results can be fully analysed.

Previous subject areas have been sugar technology, reprography, and photochemistry, each with about 1000 documents. With the photochemistry data base[6], questions were based on one known document, and answers were separately determined by exhaustive searches by other means; the aim was to investigate the operation of the indexing rules and program algorithms. A new data base, in the field of developmental psychology, is now nearing completion, with about 2000 documents. With this data base, real-life questions will be used, and it is hoped to be able to determine the effects of user characteristics and judgments on the results.

The validity of the relational system has been further proved by a program which emphasizes the relational structure. It has been found that the relations preserve meaning even when one or more of the concepts involved is uncertain or unknown, and can be represented only by an asterisk in the analysed structure of the question. For example, a search can be made for *any* compound which yields hydrogen peroxide by a given reaction. In fact, useful searches can be made by means of the relational structure almost alone, with only one known concept or possibly just an extraneous key such as a classification code number, if the data base covers a reasonably restricted field. This can overcome uncertainties of terminology.

This work has supported the validity of relational indexing as an accurate representation of the structure of thinking, and of the meaning of statements of information in documents. It requires a human indexer, but this is offset by the gain in accurate retrieval. In principle, the indexing can be made as detailed as required, and numerical and other factual data are easily incorporated. The indexing thus becomes a factual representation in which meaning is fully preserved, and the system thus gains some of the advantages of a fact-retrieval system. One is working with *information* (not just index terms), though in document-like form, and the output is similarly information, to which the document reference is only an appendage.

There are still many difficulties involved in the investigation of user effects. There may be problems of differentiating relevance from pertinence; but from the point of view of information retrieval theory, novelty of the output cannot be demanded as a criterion of efficiency, since it depends on the user's particular experience.

The great difference between a relational indexing system and the usual coordinate indexing system is that in the relational system the question remains invariant, and the indexing is varied to meet peculiarities of the user's question formulation. In coordinate indexing, the indexing remains unchanged, and the user alters his question in an attempt to obtain better results; this inevitably leads to confusion. It will be noted that the procedures in relational indexing searches do not *replace* the original indexing, but only *add* possible variants to the indexing.

Relational indexing principles thus provide a proven basis for a theory of information retrieval. This basis is found where one would expect to find it, in the nature of thinking and knowledge, at both ends of the system. Much more needs to be known, however, about the user's part in the system, before a complete theoretical picture can be developed. Finally, valid measures must be sought to express the efficiencies of the different stages, and of the system as a whole. I suggest that the examination of each stage of a system separately is the only way which offers a hope of finding such measures. It is hoped that it may then be possible to predict results and find ways of improving information retrieval. This is the ultimate goal of theory.

REFERENCES

1. J. Farradane. Knowledge, information, and information science. *J. Inform. Sci.*, 1980, 2, 75-80.
2. J. Farradane. Relational indexing. *J. Inform. Sci.*, 1980, 1, 267-276, 313-323.
3. J. Farradane and P. Gulutzan. A test of relational indexing integrity by conversion to a permuted alphabetical index. *International Classification*, 1977, 4, 20-25.
4. J. Farradane, J.M. Russell and P.A. Yates-Mercer. Problems in information retrieval: logical jumps in the expression of information. *Information Storage & Retrieval*, 1973, 9, 65-77.
5. J. Farradane. The evaluation of information retrieval systems. *J. Documentation*, 1974, 30, 195-209.
   P.A. Yates-Mercer. Relational indexing applied to the selective dissemination of information. *J. Documentation*, 1976, 32, 182-197.
6. J. Farradane and D. Thompson. The testing of relational indexing procedures by diagnostic computer programs. *J. Inform. Sci.*, 1980, 2, 285-297.