INFORMATION RETRIEVAL: THEORY AND PRACTICE

C.J. van Rijsbergen

Rapporteur: Dr. B.N. Rossiter

Introduction

If one were to use the term information storage and retrieval in a general sense then one could say that really there are three types of systems:

(1) Document Retrieval
(2) Data Base Management
(3) Question Answering.

However, traditionally information retrieval (typically abbreviated IR) has been identified with document retrieval (sometimes also known as reference retrieval). It is this first class of systems that I shall be primarily concerned with; the others will only be discussed in terms of how they relate or are distinguished from the first. Document retrieval systems are concerned with the retrieval of references to documents which will contain the information the user is seeking. For example, the request:

'Give me some references of automatic classification'

would be satisfied by the references:

'Mathematical Taxonomy' by Jardine and Sibson,
'Numerical Taxonomy' by Sneath and Sokal,
'Classification and Clustering' by Van Ryzin, etc.

since these documents do indeed contain the information the requester is seeking. In the IR jargon the documents are known as the relevant documents.

To distinguish IR work from data base system work is not easy. A data base system is also used to retrieve and suppress certain objects in response to quereies but the objects retrieved have a well-defined relationship with the query: there is no uncertainty. For example, the objects retrieved will match a query precisely in the way that an interpretation will make a statement in the Predicate Calculus true. Retrieval from a data base presupposes that one knows in advance the attributes/properties of the objects one is looking for. In IR the situation is quite different. All we know is that possibly there are some documents which are relevant to the query; there may be none. We can only guess at the attributes describing these relevant documents.

Although one talks of this notion of relevance as well-defined, it has proved almost impossible to explicate. Because of this difficulty many people (including the author) have taken an operational view of it. That is, ultimately those documents are considered relevant when the user who puts the request has decided that those are the documents he wants. A consequence of this view is that one does not attempt to construct a psycho-linguistic theory of relevance which might lead to an appropriate model for retrieval. Instead one attempts by some interactive means, or trial and error, to establish by exemplar what the likely characteristics of the relevant documents are.

This last idea is fundamental to much of the current research effort in IR, so let me elaborate on it a little. A user when asked what he is looking for can usually come up with some linguistic expression of what he wants. However, this expression generally is a very ambiguous and incomplete expression of the objects wanted. Thus, even if one knew in advance what the relevant documents were, then comparing these with the linguistic query would never lead to the discovery of what the correct computable relationship is. Obviously a user would be able to expand on the semantics of both query and documents, and produce a convincing argument about a relevance relationship but that would be uncomputable. Hence we are left with the problem: how does one guess intelligently at documents relevant to a query.

Earlier research in IR concentrated on making these guesses by partial match techniques and assumed that the more a document matched a query the more it was likely to be relevant. At the same time actual operational systems concentrated more on exact match techniques, particularly of the Boolean kind. But very early work in IR (late 50's and early 60's; IR 'started' in 1945) had discovered that one way of dealing with the inherent uncertainty associated with relevance was to model the structures and process of IR in probabilistic terms. Unfortunately this latter approach ran into computational and experimental difficulties and has only recently been picked up again to be developed into an important theoretical model for searching large files of document descriptions.

The probabilistic approach to the problem of finding a few relevant items amongst a large set of non-relevant items is not peculiar to IR alone. Other examples spring to mind; auditors searching for errors, detecting cancerous cells amongst ordinary cells, searching for precedents in case law, searching for records to deal with nuclear safety, searching historical data, and litigation support. They all have in common that the objects sought are distinguishable from those not wanted but that their characteristics/attributes/properties are not well-defined. It therefore seems natural to attempt to compute the probability with which an item might be relevant based on some information one has about the items being sought. In other words, given a query Q and

that each document in a set D is represented by a variable $\underline{x}$ then what we wish to estimate for each $\underline{x}$ is P(relevance/$\underline{x}$). I must emphasise that, although we speak of the probability of relevance given a document, we really mean given a particular description of a document. In fact from now on a document will be identified with its formal description unless the context makes it clear that we are talking about the actual document. If one were to compute this probability for every document in the set D then retrieval in order of these probabilities would seem to be the right thing to do. Of course superficially this looks fine: all we have to do is look at each document and estimate its probability of relevance. But how do we do this? If we had some psycho-linguistic model for relevance and we knew how to compare the description of a document with the description of the query then perhaps we could estimate this probability. The probability thus calculated would be in the nature of a logical probability, that is, one which is based on the comparison of propositions rather than frequencies. Unfortunately this approach, although potentially powerful, looks intractable. Instead we try to achieve an estimate through looking at the frequency of certain data items.

This last sentence disguises a lot of reduction. What I am saying is that it is possible to connect the distributions of descriptions, for example, keywords or index terms with relevance. However, I have said nothing about the way one might arrive at appropriate descriptors. In fact it is extremely difficult to separate the problem of representation from that of searching, which relies on discrimination, as I will now show.

Discrimination and/or representation

There are two conflicting ways of looking at the problem of characterising documents for retrieval. One is to characterise a document through a representation of its contents, regardless of the way in which other documents may be described; this might be called representation without discrimination. The other way is to insist that in characterising a document one is discriminating it from all, or potentially all, other documents in the collection; this we might call discrimination without representation. Naturally, neither of these extreme positions is assumed in practice, although identifying the two is useful when thinking about the problem of characterisation.

In practice one seeks some sort of optimal trade-off between representation and discrimination. 'Optimal' from the point of view of retrieving relevant documents and suppressing non-relevant ones. Traditionally this has been attempted through balancing indexing exhaustivity (the more index terms the better) against specificity (the more precise the index terms the better). Most automatic methods of indexing can be seen to be a mix of representation versus discrimination. In the simple case of removing high frequency words

by means of a 'stop' word list we are attempting to increase the level of discrimination between documents. However, it should be clear that when removing possible index terms there must come a stage when the remaining ones cannot adequately represent the contents of documents any more.

An emphasis on representation leads to what one might call a document-orientation: that is, a total preoccupation with modelling what the document is about. This approach will tend to shade into work on artificial intelligence, particularly of the kind concerned with constructing computer models of contents of any given piece of natural language text.

An emphasis on discrimination leads to a query orientation. This way of looking at things presupposes that one can predict the population of queries likely to be submitted to the IR system. In the light of data about this population of queries one can then try and characterise documents in an optimal fashion. For example, if one could estimate the probability that if a user were to submit a single-word query w he would be satisfied with document d, then comparing this probability with some user population dependent threshold could lead to an optimal indexing rule.

Probabilistic Indexing

There is a formal model of indexing which attempts to balance the importance of a term in representing the contents of a document against its importance as a discriminator. This model is based on some statistical assumptions about the distribution of words in text. One assumes that stop words are closely modelled by a Poisson distribution over all documents and that 'count-bearing' words are not. That is, a word randomly distributed according to a Poisson distribution is not informative about the document in which it occurs. At the same time the fact that a word does not follow a Poisson distribution is assumed to indicate that it conveys information as to what a document is about. This is not an unreasonable view: knowing that the word 'war' occurs in the collection one would expect it to occur only in the relatively few documents that are about 'war'. On the other hand, one would expect a typical stop word such as 'for' to be randomly distributed.

One can make the further assumption that a document can be about a word to some degree. This implies that in general a document collection can be broken up into subsets, each subset being made up of documents that are about a word to the same degree. The fundamental hypothesis made now is that a content-bearing word is a word that distinguished more than one class of documents with respect to the extent to which the topic referred to by the word is treated in the documents in each class. These content-bearing words could be mechanically detected by measuring the extent to which their distributions deviate from that expected under a Poisson process.

However, we can do better than that: the status of one of these content-bearing words within a subset of documents of the same 'aboutness' is one of non-content-bearing; that is, within the given subset it does not discriminate between further subsets. Therefore, if one assumes that there are two 'aboutness' classes then a content-bearing word w can be described by a mixture of two Poisson distributions as follows:

$$f(n) = p \frac{e^{-x} x^n}{n!} + (1 - p) \frac{e^{-y} y^n}{n!}$$

where p is the mixing probability, x and y the mean occurrences in the two classes, and f(n) the probability of w occurring n times in a document. It is important to note that f(n) describes the statistical behaviour of a content-bearning word over two classes which are about that word to different extents; these classes are not necessarily the relevant and non-relevant documents for a query consisting of that single word. When one is faced with multi-word queries it is not at all obvious how the different 'aboutness' classes relate to the set of relevant documents for the query. One needs to make some assumption about relating 'aboutness' with relevance.

Without going into details I would just like to specify the two quantities that are used in making the decision whether to assign a word w to a document as an index term or not. The first of these is the probability that a particular document belongs to the class which treats w to an average extent x ($x > y$) given that it contains exactly k occurrences of w:

$$\frac{p\, e^{-x} x^k}{p\, e^{-x} x^k + (1-p)\, e^{-y} y^k}$$

The second is a quantity involving a cost function based on the cost a user is prepared to attach to errors the system might make in discriminating relevant from non-relevant documents. If we make certain assumptions relating 'aboutness' to relevance this reduces to

$$\frac{x - y}{(x + y)^{\frac{1}{2}}} \qquad (x > y)$$

which is a measure of the divergence between the two Poisson distributions. Thus a possible measure of indexability combines the measures of representation and discrimination.

Probabilistic Retrieval

We now leave the problem of document representation and return to the problem associated with the retrieval of relevant documents given that we have settled how to describe documents and requests. For simplicity I will assume that both queries and documents will be described by the absence and presence of index terms, that is, they are represented by simple binary vectors.

Before explaining in some detail how one might define a probabilistic retrieval mechanism I shall make some assertions about the set of documents in relation to a given query. It is important to remember that throughout this section one is thinking of retrieval with regard to one typical query. Naturally the analysis will apply to any query.

There are two sets of documents with the following properties:

(1) One set is relevant and therefore wanted by the user, the other is non-relevant and not wanted by the user.

(2) These sets are in principle distinguishable.

(3) Obviously they are semantically distinguishable but we cannot compute that distinction.

(4) The description of these sets are statistically distinguishable and that this distinction can be computed.

The approach we take is to devise a mechanism which will distinguish the wanted from the unwanted documents by statistical means making as few errors as possible. Therefore a fundamental assumption we must make is that the distribution of descriptions on the relevant documents is different from the distribution of descriptions on the non-relevant documents and that this difference can be used to find relevant documents. The main quantity estimated to get at this difference is P(relevance/$\underline{x}$), that is, the probability of relevance of a document given its description $\underline{x}$. The higher the probability the more likely we are to retrieve that document. In the following it will help if the reader keeps in mind that P(non-relevance/$\underline{x}$) = 1 - P(relevance/$\underline{x}$).

The simplest retrieval rule consistent with the above considerations is undoubtedly,

P(relevance/$\underline{x}$) P(non-relevance/$\underline{x}$) $\underline{x}$ is relevant,$\underline{x}$ is non-relevant D1

(The meaning of E p,q is that if E is true then decide p, otherwise decide q.)

This is a good rule for the reason that it minimises the expected probability of misclassification. The probability of misclassification is given by,

$$P(\text{error}/\underline{x}) = \begin{cases} P(\text{relevance}/\underline{x}) & \text{if we decide } \underline{x} \text{ is non-relevant} \\ P(\text{non-relevance}/\underline{x}) & \text{if we decide } \underline{x} \text{ is relevant .} \end{cases}$$

Thus there are two types of error: one of ommission and one of commission. By following D1 we will minmise P(error/$\underline{x}$) for each $\underline{x}$. In doing so we will also minimise the expected probability of misclassification viz.

$$P(\text{error}) = \sum_{\underline{x}} P(\text{error}/\underline{x})P(\underline{x}),$$

where P($\underline{x}$) is the unconditional joint probability. A different way of specifying the retrieval rule is to rank the documents in order of their probability of relevance and to retrieve them in that order. This retrieval rule can be shown to be a good one in the same way that D1 was.

The above theory is simple and would work as if by magic if we knew how to estimate P(relevance/$\underline{x}$)! Unfortunately that is not simple. The main attack on estimating that probability is through the use of Bayes' Theorem which in this context reads

$$P(\text{relevance}/\underline{x}) = \frac{P(\underline{x}/\text{relevance})P(\text{relevance})}{P(\underline{x})}$$

where P($\underline{x}$/relevance) is the likelihood of relevance given $\underline{x}$, and P(relevance) is the prior probability of relevance. Bayes' theorem will also give an expression for P(non-relevance/$\underline{x}$). Substituting in D1 the comparison between the two probabilities reduces to P($\underline{x}$/relevance)P(relevance) > P($\underline{x}$/non-relevance)P(non-relevance), since P($\underline{x}$) is the same on both sides of the inequality and so can be ignored. Here we are comparing the probability of a description conditioned on it deriving from either the relevant or non-relevant sets. In orther words we are back to distinguishing the statistical descriptions of one set from another. If we now had some summary information about the statistical behaviour of the relevant and non-relevant documents it would enable us to estimate the probabilities for any document description $\underline{x}$. How this works in practice is explained below.

Much of the recent research work has been concerned with the assumptions that can be made about the form of P($\underline{x}$/relevance) and P($\underline{x}$/non-relevance). If as stated earlier we assume that $\underline{x} = (x_1, \ldots . x_n)$, a binary vector reflecting absence (x_i=0)

or presence ($x_i=1$) of each index term i from 1 to n, then one could assume the index terms to be independently distributed. This simplifies the decision rule considerably. In terms of the binary document space it amounts to constructing a linear decision surface which classifies the points on one side as relevant and the points on the other side as non-relevant. More elaborate and more realistic assumptions such as assuming certain dependencies between index terms will lead to non-linear decision surfaces.

It must be stressed that although these theoretical developments are elegant and promise effective retrieval in terms of making fewer errors, there are many problems associated with them that remain unresolved. The most important one is probably that associated with making the actual estimates for the probabilities. The way this works in practice is that one uses a simple, fast, crude retrieval strategy to retrieve some relevant documents. From this retrieved set one then estimates the statistical properties of the relevant and non-relevant documents. These sets are intended to be very small since a user must decide on the relevance or non-relevance of each document in the retrieved set. This means that small sampling theory must be invoked but this is not valid because the retrieved set is not a random sample. Another problem is associated with the dimensionality of the space in which the documents and queries are represented. There appears to be an optimal dimension beyond which the errors incurred by rule D1 increase!

Different Approaches

One important thing to note about the above approach to document retrieval is that to operate the model one needs to acquire some knowledge of the relevance or non-relevance of a small number of documents. There are strategies which do not require this kind of knowledge. Instead one builds structures which are of some help in guiding the search for relevant documents. Interesting structures for this purpose are of a classificatory nature. I believe it to be fundamental to the process of finding the so-called relevant documents that one uses the classificatory structures that underlie the different types of items of information that are stored. To put it differently, it is precisely the classifications inherent in the data that will help us find the relevant documents.

Two classificatory structures that have receved much attention in IR are generated by document clustering and index term clustering.

In document clustering one is concerned with the automatic classification of documents for the purpose of providing more effective and efficient access to them. That it is likely to provide more efficient access is not difficult to see. By grouping the documents appropriately, one will be able to limit the search for relevant documents to only a small part of the document collection. In principle this sounds fine. We use the classificatory structure

(just like the U.D.C.) to guide us to the chunk of the collection containing the relevant documents. That this can be done efficiently is not difficult to see if one has a hierarchic classification of the documents. By starting the search at the top level one can eliminate an ever increasing proportion of the collection not likely to contain relevant documents. In saying this, we are assuming that the search strategy based on clustering of the documents can actually find a large proportion of the relevant documents. In other words, to speak of efficiency only makes sense in the light of an effectiveness criterion; it is not difficult to design a highly efficient search strategy that will find nothing. The claim is that by clustering the documents we can achieve a certain level of effectiveness more efficiently than by other methods not using clustering. Experiments on a variety of document collections and with a variety of clustering methods have shown that, in principle at least, this claim can be met. A more ambitious claim is that document clustering can do better than strategies not using any information about the relationships between documents. This claim is much harder to justify: there is some theoretical evidence but experimental evidence is sadly lacking.

In term clustering one is attempting to construct a structure that relates index terms in some useful way. A typical example would be the construction of a thesaurus by some clustering method. Term clustering in information retrieval is generally used either to modify a request for documents or to alter the document descriptions in the collection so as to increase the probability of finding the documents relevant to a user's request by a search strategy. The data used to construct the clusters of terms is mostly derived from the relative frequency with which index terms co-occur in the total document collection. In other words terms which co-occur relatively frequently will form tight clusters, whereas a pair of terms that do not co-occur to any extent will find themselves split between clusters. The underlying assumption for doing this is that reference to one index term either in a request or a document should automatically lead to the consideration of the closely associated index terms.

A Probabilistic Basis for Clustering

As noted earlier, to construct a document or term clustering no prior knowledge about the relevance or non-relevance of documents to the population of potential queries is used. But when we use these classifications in retrieval we expect that they will provide us with a guide to the relevant documents for any query whatsoever. So somehow a connection between a classification of documents or index terms and relevance is made. Since a classification can be derived from an association measure, a connection is, therefore, also made between association of documents or index terms and relevance. An attempt to clarify this connection iis embodied in two hypotheses, one the Cluster Hypothesis for documents, and the other the Association Hypothesis for index terms. These hypotheses will now be discussed in a little more detail.

The Cluster Hypothesis states that closely associated documents will tend to be relevant to the same requests. It forms the basis for document clustering. As it stands it is a fairly weak statement and it is possible to tighten it up. One way of doing this is to consider the probabilistic version of it which might read as follows: If document x is closely associated with y, then over the population of potential queries the probability of relevance for x will be approximately the same as the probability of relevance for y. In symbols P(relevance/x) and P(relevance/y) will be of comparable size for any given query. This tighter formulation certainly implies the Cluster Hypothesis.

The second hypothesis, the Association Hypothesis, says that, if one index term is good at discriminating relevant from non-relevant documents, then any closely associated index term is also likely to be good at this. This hypothesis is trivially true for index terms that occur in the same set of documents. However it is the intermediate cases that are of interest, namely those where one term occurs in one set and another term occurs in an overlapping set. In such a situation the hypothesis makes a strong claim about the effectiveness of closely related index terms. It tells us exactly how a term classification should be used in a retrieval operation. After one has discovered the effectiveness of a particular query term submitted by the user, the class mates of that term are also likely to be effective retrieval terms.

For each of these classificatory structures, documents or term, it is clear how it should be used on its own in retrieval. Some tentative attempts have been made to devise a single structure containing both the term-term and document-document relationships so that both classifications could, in principle, be derived from this single structure. This is an appealing idea but one that has not got very far yet. It may be that, by looking for a suitable hypothesis that will imply both the Cluster and Association Hypotheses, we will be able to find a suitable structure to exploit both the association between index terms and documents.

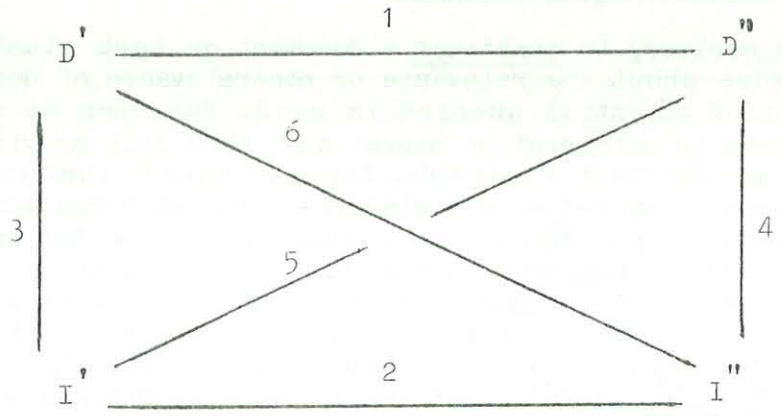

Figure 1

A Unified Theory?

It seems that the models of IR incorporating clustering all centre around the diagram in Figure 1. Let D' and D" be any two documents and I' and I" any two index terms. When we are clustering either the documents or the index terms we calculate the strength of association implied by link 1 or 2. For example, the strength of link 1 depends on the extent to which the index term assignment to D' is similar to the assignment to D". The usefulness of this link is them implied by the Cluster Hypothesis. Similarly the usefulness of link 2 is given by the Association Hypothesis. To exploit these document-document and term-term links, we generate classifications so that only the important ones are represented. Now let us look at the other links in the diagram. The strength of one of these is a function of the context in which the linked items occur. For example, the strength of link 6 between D' and I" may depend on the frequency of occurrence of I", or depend on any other contextual information that is appropriate. Just as in document-document or term-term associations some of these document-term links will be more important than others. I conjecture that there is a reasonable structure (and corresponding hypothesis), different from a standard classification, underlying the use of such diagrams as in Figure 1 which will incorporate the significant associations. Given such a structure, the remaining difficulty will be to try and connect it with relevance in the way that the Cluster and Association Hypotheses have done for classifications.

A small step in that direction is given by the Discrimination Gain Hypothesis, which reads as follows:

> Under the hypothesis of conditional independence the statistical information contained in one index term about another is less than the information contained in either index term about relevance.

To understand this hypothesis some comments are in order. The 'conditional' independence here refers to the statistical independence of the index terms on both the relevant and non-relevant sets. 'Information' is used in the strict sense of the expected mutual information or Shannon's channel capacity. And of course one assumes that there is an underlying, hidden, binary variable, relevance, about which one is trying to get information. With some parametric assumptions about the underlying statistical distributions one can in fact prove the result, but unfortunately I am still unable to find an unparametric proof.

The way this hypothesis would be used is to steer the search from term to term, deciding in the light of the strength of the term-term links which terms are likely to lead to relevant documents. Of course this is only part of the story; to complete it one would have to make statements about the relationships of the other links

with relevance. For example it may be possible to relate the term-document links to relevance through the probabilistic theory of indexing described above. If this were done then at any stage of the search one would be able to decide whether it wwas more profitable to look at a connected document or whether to look at a connected term.

Conclusions

In this paper I have concentrated on presenting some of the more important recent developments in IR with here and there some comments about practical implementation. This I have done deliberately for the important reason that the implementation of research ideas in IR has run into practical difficulties. The hope is that some of the recent developments in data base management systems will alleviate some of the practical problems we have in IR.

In the past most experimental work in IR has been on a small scale and special purpose software has not been too difficult to write. However, the impact of these experiements has been rather slight. For whatever reason that may be, the current demand is for large scale controlled experiments in IR, for example, testing theories of probabilistic indexing and retrieval. To perform these tests would require a massive investment in special purpose software. Instead I think the right way to go is place piggy-back IR systems on some existing data base management system (DBMS). This is not an easy matter. Most DBMS's are not designed to handle IR queries in quite the form that is required, although the recent attention paid to natural language queries by some researchers in the data base area may change that. A further problem is that, for efficiency reasons, text storage in IR must be done through indirection otherwise matching will become incredibly slow, apart from the increase in storage entailed by the direct storage of text strings. Some of the initial text processing required before document descriptions are entered into the data base is also difficult within the DBMS: for example, stemming and conflation.

Many of the structures generated for the purpose of aiding retrieval such as terms and document classifications are easily set up as relations. Thus an obvious candidate for implementing an IR system incorporating those structures would be a relational data base. Unfortunately the relational algebra or calculus is not immediately suitable for expressing IR queries so that a front-end processor translating IR queries into the data base query language will have to be designed. This translator need not necessarily be very complex unless one wishes to take into acccount much of the syntax of the query.

The retrieval problems associated with IR are somewhat different from the standard ones tackled by DBMS's. In commercially available retrieval systems one can retrieve on the contents as well as on such attributes as author, journal, cited author, etc. So the pay-off for

using a flexible database management system in IR could well be quite large since some of the operations required are easily executed by a database system. It remains to be shown whether an existing database system with possibly some modifications can provide the basis for an efficient, powerful, and flexible IR system. Powerful, in that it implements some of the theories presented in this paper, flexible, in that it will retrieve both on contents and bibliographic keys.

Select Bibliography

ATKINSON, M.P., 'Database systems', Jorunal of Documentation, 35, 49-91 (1979)

HARTER, S.P., 'A probabilistic approach to automatic keyword indexing, Part 1: On the distribution of speciality words in a technical literature, Part 2: An algorithm for probabilistic indexing' Journal of the American Society for Information Science, 26, 197-206 and 280-289 (1975)

MARON, M.E., (Ed.) 'Theory and foundations of information retrieval', Special Issue, Drexel Library Quarterly, 14 (1978)

ODDY, R.N. 'Reference retrieval based on user induced dynamic clustering', Ph.D. Thesis, University of Newcastle upon Tyne (1974)

ROBERTSON, S.E., 'Theories and models in information retrieval', Journal of Documentation, 33, 126-148 (1977)

SALTON, G., 'Mathematics and information retrieval', Journal of Docuementation, 35, 1-29 (1979)

VAN RIJSBERGEN, C.J. Information Retrieval, Second Edition, Butterworths, London (1979)

Discussion

Professor Randell asked (1) whether information retrieval techniques had been applied in practice to areas other than documents and (2) whether people working with database management systems were aware of past work in information retrieval? In reply to the first question, Dr. van Rijsbergen replied that they had been used using pattern-matching techniques in cancer research and in nuclear safety. In reply to the second question, Professor Tsichritzis thought it was sometimes easier to reproduce what had been done before rather

than to patiently search the literature. It was important to note as well the important technical differences between information retrieval and database management systems. The former provided no update facilities and handled varriable-length data which was often textual and unstructured. The latter provided comprehensive update facilities and handled fixed-length data which was highly structured. Only if a structure could be found in textual data, would it be suitable for database management systems.