4

Ineffable concepts in information retrieval

Nicholas J. Belkin

4.1 Introduction

There are a number of concepts (phenomena, entities) which make the testing of information retrieval systems especially difficult. The difficulties they pose arise primarily from their elusive yet ubiquitous nature, since they are difficult to define, either operationally or conceptually, yet they appear central to the information retrieval situation. In my view, that situation is dependent upon the problem of information science, which can be stated as:

> the effective transfer of desired information from human generator to human user[1].

This problem is further specified by the information retrieval situation which can be characterized as follows:

(a) A user, recognizing an information need, presents to an information retrieval mechanism (i.e. a collection of texts, with a set of associated activities) a request, based upon that need, hoping that the information retrieval mechanism will be able to satisfy the need.
(b) The task of the information retrieval mechanism is to present the user with the text (or texts) which it judges to be most likely to satisfy the user's information need, based upon the request put to the mechanism.
(c) The user examines the text, or some or all of the texts, presented by the mechanism, and her/his need is satisfied completely or partially or not at all. The user's judgement as to the contribution of each text in satisfying the need establishes the **usefulness** or **relevance** of that text to the need.

Thus the fundamental issues with which information retrieval is concerned must include:

> information need; desire; information; aboutness or meaning; satisfaction (including relevance); and effectiveness (of information).

These issues or concepts attain their importance because of the way in which information retrieval systems attempt to solve the problem of information science, and in that context can be roughly categorized as follows.

(1) User-dependent concepts: information need; desire.
(2) Text-dependent concepts: information; aboutness; meaning.
(3) User and text-confounded concepts: satisfaction; effectiveness; synthema (homeosemy).

This categorization of issues derives from the general structure of information retrieval systems[2], in which documents and needs are each separately represented, then matched against one another in order to retrieve documents which are judged by the user according to their appropriateness to her/his need. This situation requires concepts basic to need representation and need understanding, concepts basic to text representation and understanding, and concepts concerned with the relationships between text and need. Various concepts basic to each of the three areas outlined above have been widely discussed in the literature, although not always to great effect as far as testing of information retrieval systems is concerned. Usually, the test of any system has been concerned with secondary entities or processes in one of the sub-areas (such as comparative experiments on indexing systems), stopping short of investigating the relationship of underlying concepts (such as **information** or **aboutness**) to the results of the tests, or even of determining whether there were any such underlying concepts to the systems being tested. The suggestion here is that it may now be the right time to begin such investigations, to make these concepts at least explicit in testing, and perhaps even to make them the basic variables in the testing of information retrieval systems. Before continuing this argument, some general discussion of these concepts themselves is in order.

4.2 Definitions or interpretations of the concepts

User-related concepts

Although this group of concepts seems to be the obviously central core to the information retrieval situation, since evaluation of system performance should be solely in its terms[3,4], it seems to be the most neglected in the literature of information retrieval system testing. This may be because concepts such as *relevance*, which depend upon this group but are confounded with the text-related concepts, have been initially more important to systems testers in that they provide the means for direct comment on system performance. There has been somewhat more treatment of user-related concepts in such areas as reference work[5] and in theoretical discussions of information science[4].

The basic situation, as outlined by Taylor[5], is that of a person coming to an information system with some already (at least vaguely) recognized need, and going through various stages of representation of that need which culminate in a formal request put to the information retrieval mechanism in terms which it can use for matching against its store of texts. In this situation, one can recognize a number of elements which are likely to affect the mechanism in significant ways, yet which are difficult to describe or quantify.

The first of these is the **desire** of the user. This concept seems not to have been discussed explicitly as a separate issue, but one should note that, apart from a **need** for information, the user comes to the mechanism with some set

of constraints on what will be acceptable or reasonable replies to the expression of need. These constraints are related to the reasons for wanting to satisfy an information need at all, that is the user's desire. Wilson and Streatfield[6], for instance, have suggested that there can be significant affective reasons (e.g. wanting to keep up with or ahead of subordinates) for obtaining information, which have been relatively little studied in information science. In any event, it seems clear that, no matter what the information need in a conceptual sense, the context of the need will be important in judgements of the extent to which the information retrieval mechanism has satisfied the need, and in determining the mechanism's response.

Aspects of this context have been described by Wersig[7] as the **problematic situation**. That is, the user's conception (or model) of a real-life situation which the user has recognized as being in some way inadequate. The desire here is to acquire the information necessary to resolve the problems in the model. Belkin and colleagues[8,9] have suggested that that which underlies the information need is an **anomalous state of knowledge**—the recognition by the user that her/his knowledge of a topic or situation of concern is inadequate. Both of these ideas are attempts to make the concept of **information need** more amenable to formal manipulation.

The information need leads to a **question** or **request**. Although there has been some investigation of the relationships between needs and questions, and of the nature of questions in the information retrieval literature[5,10,11], these too appear to be still rather inadequately discussed issues. The tendency in information retrieval testing has been to accept that there is a difference between need and question, but then to deal with questions rather than needs. Obviously, evaluating a system's performance must depend, at some stage, on making the relationships somehow explicit, but this, as we will see later, can cause grave difficulties.

Text-related concepts

This cluster of concepts has been far more extensively discussed than the user-related group in the information retrieval literature. An obvious reason is that information retrieval systems design has typically been concerned primarily with text representation issues, another that it somehow seems intuitively reasonable that *information* is what this science is all about, and that information somehow is related to or inheres in the text. Belkin[12] has reviewed a number of information concepts proposed for information science. The general problem with most is that of reconciling a desire for predictability with the observation that the same text will affect different people differently. Some have attempted to define out the individual variability by considering only the text[13]; others have given up predictiveness in favour of completely individual-based notions of information[14]. The controversy is not resolved, nor does it appear likely to be, for we deal here with a **concept**, which one determines according to the needs of a situation, rather than a definition of a **phenomenon** applicable to a wide range of situations or contexts. Nevertheless, most text representation schemes depend upon some (usually implicit or inexpressed) concept of information, which generally assumes that information is a quality or quantity associated

with a text. This makes ferreting out that concept, or making it explicit, of potential importance in evaluation of such schemes.

The concepts of **meaning** and *aboutness* are also central in the question of text representation, presenting similar difficulties of interpretation. In general, information retrieval has been concerned with what a text is **about** (i.e. some topic specification) rather than what it **means** (e.g. the propositional structure of the text). But just what the relationship between these two concepts is, and how one interprets them, is still a significant problem. Some describe aboutness as a whole-text phenomenon[15], others as a phenomenon related to the state of knowledge of the reader[16]. The two positions are quite different in their predictions and in their prescriptions concerning text description systems, which makes it important when evaluating them to know on what premises the system has been based. Robertson[17] and Sparck Jones[18] have recently provided good discussions of the relationship between meaning and aboutness in the information retrieval context. The major effect that such problems might have on information retrieval system testing is that different representation systems might be based on different concepts of aboutness, or might even be based on a concept of meaning (as, for instance, one assumes the LEADERMART system is[19]), and thus differences in performance might be explainable only in terms of these underlying concepts. Certainly, as Gardin[20] demonstrates, systems based on meaning will be substantially different from those based on aboutness.

Confounded concepts

This group of concepts is concerned with the relationship between need and text, and can be most generally described as having to do with 'satisfaction' (with two major exceptions). This group, or at least some aspects of it, is the one of the three discussed which has been most thoroughly investigated in the context of information retrieval system testing, for the good reason that it provides the basis for evaluation of information retrieval system performance.

The best known of these concepts is *relevance*. Relevance, in its most general information retrieval sense, describes the appropriateness of a text to a specific information need. Saracevic[21] has written a fine review of research and speculation on this concept in information science, but we can mention here a few of its problems and a few proposals for dealing with them. A major difficulty in information retrieval system testing has been to obtain reliable and reproducible relevance judgements. The problem has been that the ultimate judge of relevance in the real world is the person with the information need, and this person's formal question to the information retrieval mechanism (which is all that mechanism has to work with) may not encompass all of the factors which bear upon her/his eventual relevance judgement. So relevance as a general concept has tended to be divided into two or three separate concepts (see, e.g. Kemp[22] or Wilson[23]). These include: **logical relevance**[24], which requires that the propositions of the request be included in, or logically deducible from the text; **destination's relevance** or just **relevance**, which depends only upon the relationship between the topics of the formal question and the texts retrieved, usually as evaluated by an external judge, and **situational relevance** or **pertinence**, which is the judgement

by the user of the appropriateness of text to need based not upon the question put to the system but rather on the user's entire desire and need state at the time of receiving the text for judgement. Each of these concepts of relevance is obviously quite different from the others, and each has special advantages and problems in testing situations. The major point here is that the choice of relevance concept will radically affect evaluation of system performance.

The most commonly used measures of performance, recall and precision, depend solely on *relevance* judgements, and thus are peculiarly sensitive to this concept. **Utility**, the other major proposal for performance evaluation (see, e.g. Cooper[25]), can in principle take account of other factors than relevance in the destination's or logical senses, but at the moment practical difficulties in assessing this measure make its use difficult.

Karlgren[26], with **homeosemy**, and Robertson[27], with **synthema**, have independently attempted to formalize the topic relationship between text and question. They both suggest that the problem may be considered as the degree to which the aboutness of each coincides with that of the other, and that this relationship, especially in Robertson's suggestion, might provide a scale which would underlie the concept of relevance. Notice the extent to which both of these notions depend upon the concepts of aboutness or meaning. These suggestions provide a potentially valuable way of partitioning the various aspects of the relevance concept, but as yet are still only suggestions, with no concrete methods of implementation.

To return to the general notion of **satisfaction**, one can see that logical and destination's relevance, and more formally synthema or homeosemy, refer only to conceptual aspects of the need, and assume that the question is an adequate representation of the need. Situational relevance extends the notion to include beliefs and other aspects of the user's condition, while accepting the possibility of adequate linguistic representation, but pertinence and utility go even farther in attempting to take into account affective and other aspects of the need, and do not necessarily assume that the question expresses the need in its entirety. *Usefulness* is another term that has been used in place of relevance in attempts to include more than strictly topic-related aspects of need in the judgement process. Apart from satisfaction, in terms of relevance or utility, **effectiveness** appears to be the only major candidate as a basis for information retrieval system evaluation. If, for example, information is considered as 'data of value in decision making[28]', then system performance can be evaluated by whether its output was used in making appropriate decisions; that is, by some 'objective' or at least behavioural measure of the effect of its output on the user. This approach completely bypasses the question of relevance, but is unfortunately very difficult to operationalize, especially in obtaining and measuring observations of the effect. It also poses significant difficulties in prediction of effect, because of its situation dependence.

4.3 Interactions

The concepts discussed in this chapter are difficult to deal with in an experimental or investigatory setting not only because they are conceptually and practically not very tractable, but also because they are highly

interdependent. Whether one wishes to design experiments to study one of the concepts as such, or to be able to control for their effects in a system test, or to use one or more of them as an explanation for system performance, sorting them out from one another always remains a difficulty. For instance, the concepts of information need and of desire always must underlie any satisfaction concept. Thus, any discussion of relevance depends upon some idea of information need. This means that, in any test, the relationship of the relevance concept being used to the information need concept underlying it ought to be made explicit, and the implications of the relationship should be at least mentioned. The *questions* used in information retrieval tests also depend on the concept of need, usually assuming only topic-related issues. But question representation also depends upon the system used for text representation, which in turn depends upon concepts of information or aboutness. And in those cases in which the question is used as the basis for judging relevance, relevance becomes dependent upon the question concept, and the circle is completed.

It is obviously necessary to eliminate or control for these interactions in testing situations, and this is indeed possible, but with some attendant costs. For instance, in the Cranfield experiments the questions were purposefully composed without any underlying information need, thereby allowing the experimenters to use a relevance concept based solely on topic relations between question and text. In this way, the systems for text and question representation could be evaluated without reference to information need or desire, the user-related concepts. The cost for having been able to disentangle these concepts lies in the assumptions that questions without needs are the same as questions with needs and that performance judged by topic-based relevance is at least directly related to performance judged by need-based relevance.

In experiments designed to investigate one of these variables *per se*, the problems of interdependence are perhaps somewhat easier to control for. But, as always, questions of relevance must depend upon concepts of need and of information or aboutness, and empirical investigations of information are difficult to separate from the states of knowledge of the subjects receiving the information. There are, unfortunately, no good and general rules or techniques for isolating one of these variables in any given testing situation. The usual solutions have been to hold as many variables as possible as constant as possible, or to establish control groups for variables that cannot be held constant. Usually, these designs require strong assumptions about the nature of the interactions among the variables, and about the variables themselves. The most general suggestion that one can make in these cases is that the best way to discover the least obtrusive and confining design is to make these basic assumptions explicit, and on their basis to establish the control structure required. The next section will discuss some possible operational definitions for these ineffable concepts, and also their problems.

4.4 Inference chains and operational definitions

A fundamental difficulty with these concepts is that they are very basic indeed. This means that theories about them are often very general, and that

it is difficult to make explicit predictions of behaviour or other empirically verifiable phenomena on their basis. And, for the same reasons, it is very difficult to determine reasonable operational definitions for these variables. In order to achieve these goals, it is usually necessary to go through a number of subsequent assumptions or hypotheses, each of which is a theoretical construct in its own right. When one finally gets to some phenomenon that is operationally definable or empirically observable, the relationship of that phenomenon to the original theoretical concept is probably very tenuous indeed. All of the intervening constructs and assumptions mean that it is unclear just what is being tested in the final experiment or investigation. Concepts from both the user and text related groups share this problem, and so, therefore do those from the group of concepts arising from their relationships.

For example, consider the problem of operationalizing information need. Belkin and Oddy[9] have suggested that an 'anomalous state of knowledge' (ASK) is the basis of any information need, and that information retrieval systems should attempt to use representations of ASKs as the basis for retrieval. An ASK is considered by them as a part of an individual's state of knowledge which that person considers to be inadequate (anomalous) in some way. The first problem that arises in trying to make this concept operational is to decide upon a general schema for representation. On the basis of psychological arguments, the investigators[29] chose structures consisting of concepts and relations among the concepts. Next one needs to decide upon means for obtaining the data from which the representation will be constructed. They decided to use 'problem statements'; that is, statements by users about the problem which brought them to an information retrieval system. This decision was supported by Wersig's[7] argument concerning the problematic situation, but the method for eliciting these statements had to be designed from first principles. Finally, a technique for analysing the data and generating the structure is needed. On the basis of some quite speculative argument concerning underlying 'cognitive' structures and their reflection in linguistic structures, and in order to make the problem relatively simple, the general structure chosen was one of associative relations among concepts, these concepts to be represented by word stems and strength of association determined by the degree of co-occurrence of words within specified distances in the text of the problem statement. This entire chain then resulted in a structure which was claimed to be a representation, at some level, of the ASK underlying the person's information need. The representation could be displayed as a graph, with word stems as nodes, associative relations between nodes represented by edges, and the distances between nodes related to the strength of their association.

Consider now what lies between the original theoretical construct (the notion of an ASK) and its operational definition. There are assumptions and decisions made about what a state of knowledge is, or could be; about how, and even whether, some verbal description of an ASK can be elicited; about the nature of relations between concepts in a state of knowledge; about the relationship between the distance between words in a text and association strength of concepts in a state of knowledge; and many more. These assumptions build one upon the other in an elaborate inference chain, so that the end product, the representation, is only tenuously related, and in very

uncertain ways, to the original ASK idea. The problem now is: how can one test the validity or accuracy of that original construct? For instance, if one wishes to know if the end representation is an 'accurate' reflection of the user's ASK, and tries to determine this by direct question, the response may be dependent upon any one of the assumptions in the entire chain, with no easy way to tell which of them, or which combination, is at issue in the response.

In principle, one ought to examine and test the validity of each of the hypotheses made along the way before performing the test as a whole, but in practice this is very unlikely to happen. The more generally appropriate strategy is to attempt to develop a design in which, as in this example, responses can be directed to each confounding factor. With a chain of assumptions the length of this example, which is probably not unusual, the design may become very complex, and the test instrument clumsy. A possible approach is to run a series of tests, each concentrating upon the end product from the point of view of one of the assumptions, using the data derived from each of the series for design of the subsequent members. But in such a case, there will be some assumption or hypothesis that one cannot, or will not be willing to test (in the example, perhaps the idea of knowledge as a structure of concepts and relations), which must then be considered as an integral part of the original construct itself.

An example of this type of problem from the text-related concept group is the question of relative **informativeness** of text representations. This is an obviously important issue in comparative evaluation of techniques for text representation, especially if the text representations are to be used as the basis for relevance judgements or for matching for retrieval. In the formal case, the situation is that the user, or the user's representative, is presented with some description of a text, on the basis of which a probabilistic relevance decision must be made. In such a case, the obvious strategy is to compare judgements of the representations with relevance judgements of the entire documents, within each system, and then to compare the overall results of the competing systems. Belzer[30], for instance, has done an experiment of this sort. In this case, informativeness is operationally defined as the capability of the representation to induce a 'correct' relevance judgement in the user. There will be problems in such a design with possible interactive effects of the documents and document representations upon relevance judgements made by any individual, but in general the dependent variable can be fairly well isolated from problems of inference chains, as long as the test is only evaluative or comparative. But, if the purpose of the test is explanatory; that is, if one wishes to explain the differences in informativeness between the representations, then concepts of aboutness, meaning and information become important, and the inference chains from the underlying theory to the eventual representation can cause problems. One now must begin to investigate the assumptions, to see how they have influenced the representations, in order to decide whether the intermediate assumptions, or the underlying concept, are the reasons for the performance.

In the case of informativeness as applied to matching for retrieval, the situation is much more complicated immediately, for it must require an aboutness or meaning concept for its implementation. That is, informativeness here means the ability of the representation system to represent both

documents and needs so as to maximize the matching mechanism's ability to predict the topic relationship between text and need. This assumes that informativeness is an appropriate quality for accomplishing the goal, and furthermore that informativeness is dependent upon aboutness or meaning (topic), while not being identical with either. To test relative informativeness in such a context, it is no longer enough simply to accept relevance judgements, for there is not necessarily a strict correlation between relevance judgements and topic relations. It might, indeed, be possible to use independent evaluations of text and/or need topic as the basis for a test design in this context, but in order to do this properly, the idea of aboutness or meaning which underlies the informativeness notion must be used as the basis for these assignments. Note, however, that the topic assignment must be in some terms other than those of the representational scheme(s) being investigated. The following example discusses some of the inference or interpretation problems that arise in this situation in more detail.

This example of a chain concerns the notion of aboutness as applied to both text and need; that is, synthema or homeosemy. Here we are concerned with the general case of developing or testing a retrieval mechanism based on the degree of synthema or homeosemy between text and need. In order to do this, one must first begin with some notion of aboutness; say Hutchins'[15] idea that it inheres in the thematic structure of the document as a whole. From this basic idea, one then needs to develop an analytical technique for obtaining a representation of aboutness from the document structure. This technique will have its theoretical basis in text-linguistics, and will indicate the significant concepts of the document and their interrelations (say). One could, perhaps, use the resulting structure directly for matching purposes, or reduce it to, say, a set of index terms. Such reduction would again be based upon an assumption that aboutness can be adequately represented by a set of single concepts. So here is an aboutness representation of the document, which one wishes to match against an aboutness representation of a need.

Notice how many assumptions have been made here. More are needed when one comes to the information need representation. Thus the first assumption concerning the need must be that what the need is about is indeed capable of being precisely expressed linguistically. This assumption leads one to a technique for eliciting a statement of need from an information retrieval system user, which can be analysed and represented by the techniques used to analyse and represent the document (or at least by techniques that result in similar structures). These steps assume that documents and questions (linguistic need representations) are basically similar in their aboutness structures. Given this assumption, one then matches the two representations against one another, in order to judge their 'likeness'.

The question of likeness then introduces the need for a whole new set of assumptions, concerning the scale along which likeness will be determined. One solution in information retrieval has been to accept indexing-type representations, and then to assume that the degree of synthema is related to the overlap of index terms between the two representations (level of co-ordination). Other solutions include spatial or vector analogies, in which the distance between two points in a space, or the angle between two vectors[31, 32] is a measure of the likeness of the document and need represented by the two entities in the space. Notice that any of these solutions requires strong

assumptions about the independence of index terms, and about the nature of the space or scale in which the entities are to be compared.

Only after all of these three types of assumptions (text-related, need-related and matching-related) have been granted, does one actually achieve the original goal; to establish some operational means of determining the homeosemy of a document with a need. And just as in the ASK example, each of these assumptions has strong theoretical implications which ought, in principle, to be tested.

The point of this discussion has not been to discourage investigation of these complex and basic concepts, but rather to indicate the sorts of difficulties one can expect in trying to deal with them, and to make some tentative suggestions about how one might deal with them. These suggestions can be summarized as: first, make certain that all of your assumptions or hypotheses have been made explicit; secondly, try to minimize the steps in the chain from theoretical construct to operational definition; thirdly, design the test to investigate, as much as possible, the effect of each assumption; and finally, be explicit in reporting the decisions about those assumptions which were left untested. With concepts such as need, information and satisfaction, such chains will always be necessary. This does not mean that the concepts cannot be studied and included in tests, but it does mean that such tests must be unusually self-conscious in their design.

There still remains the problem of generality of theory. In the first example of this section, there was a minimal theory that ASKs underlie information needs, and furthermore that ASKs can be represented as certain types of structures. Now from such theoretical statements one can indeed generate some predictions or procedures for making them operational, but it is quite difficult to construct these so that their inadequacies or failures can be interpreted as invalidations or falsifications of the theory itself. In the ASK example, people might be asked to comment upon the relative accuracy of the representation of their need, but negative comments might have no bearing upon whether the original theory is valid, for the elicitation technique or the specific representational format might be equally at fault. Similarly, one might be able to predict that certain types of 'anomalies' might be associated with certain classes of information need, but if the prediction fails, it can be interpreted as doing so only at the representational or classificatory levels. There seems no simple way to avoid this sort of problem with these concepts, so that perhaps what one should do is to accept it, and to consider these concepts as basic assumptions which lead to particular strategies or systems for solving certain operational problems. In this way, one can evaluate a system as a whole (but taking account of inference chains) according to how well it solves the problems (or achieves its goals). This could be done in comparative or single evaluative contexts, and one would attempt to judge the theory not according to absolute validity, but rather according to how well the framework which it establishes works in the context of the problems it has been constructed to solve.

4.5 The significance of 'ineffable' concepts in information retrieval testing

There are two ways in which the variables discussed in this chapter are important in information retrieval testing: the first is as objects of study in

their own right; the second is as possible sources of error or variation in other aspects of the system with which the tester is primarily concerned. Each of these situations raises different questions and different problems of interpretation and control. The major problem in the first case lies in the construction of operational definitions of the variables to be studied. This may be difficult, but once these definitions have been established, it may be possible to study the variables in isolation. For instance, if one wants to investigate the concept of 'aboutness' experimentally, one could control the experiment to include, say, only documents and readers of documents, without reference to an information retrieval system at all. The problems of the second case, on the other hand, arise because the interactions of variables within information retrieval systems are so complex. In this situation, it is in general not possible to isolate variables completely. For instance, if one wants to evaluate some method of content description within an information retrieval system, then the relationships between the concept of aboutness, the description method and the evaluation measures are significant, even though aboutness is not an explicit variable in the experimental paradigm. Thus, in the second case, which is likely to be the usual situation in information retrieval system testing, one ought to begin by attempting to isolate all of these variables (conceptually), and then to see to what extent they actually might influence the variables in which one is interested, and the evaluation measures. In this case, again, one is not interested in the 'ineffable' variable itself, but rather in the effects that not having included it in the design might have on the results of the test. It is useful here to discuss, in rather general terms, the extent to which being able to deal with these concepts might affect an information retrieval test.

Text-related concepts

It is possible that having a specific, operationally definable and experimentally tractable information concept is a necessity for the development of theory in information retrieval. Nevertheless, there is some question as to whether this particular concept need be made explicit in evaluation tests, at least of existing systems. The reason is that there is usually only a minimal relationship between whatever passed for an information concept in the system as originally formed and the measures used to evaluate the system's performance. On the other hand, if one wishes not only to compare, but also to explain differences between systems, then both information and aboutness become quite important. This is especially true if the focus of attention is the description mechanism, rather than, say, the retrieval strategy. If one wishes seriously to explain the difference between two description mechanisms, then one must be able to discuss the relationship of each mechanism to an underlying information or aboutness concept. For instance, one could compare two systems of automatic indexing in terms of retrieval performance (recall and precision, say) and discover that one of the systems gave consistently better results. But to be able to say why this was the case, one would need to consider the underlying assumptions of each method, in terms of an underlying information or aboutness concept. One could then make some decisions about whether the difference in performance lay in a different technique deriving from the same assumptions, or in different assumptions.

In the former case, the two methods could then be strictly compared or evaluated against one another, one method definitely being said to be better than the other. In the latter case, however, the decision may not be so clear, for the difference in techniques used may not be as material as the difference in basic assumptions.

User-related concepts

This group of concepts is much more obviously likely to affect system evaluation directly than that previously discussed. The basic reason is that all evaluation measures, save perhaps effectiveness, depend strictly upon at least one of them, and in a much more obvious way than upon the text-related concepts. Therefore, it is necessary to have a well-defined concept of information need in order to be able to interpret and use properly the user's judgements of the system's performance; that is, the user's satisfaction or dissatisfaction. The Cranfield experiments, and others, recognized and attempted to control for the need problem by eliminating it entirely through the use of artificial questions (that is, questions without underlying needs). Then the relevance judgements were carried out in an 'objective' manner, untainted by individual differences among variable users. This strategy is useful in that it explicitly recognized the difficulty of dealing with individual information needs. The problem is that there is no *a priori* reason to suppose that the performance of a system measured in this way correlates at all well with performance as evaluated by posers of real questions.

Furthermore, such evaluation techniques tend to assume that the user needs or desires all of the relevant documents. Cooper[33] and Oddy[34], among others, argue cogently against this assumption, and it seems that in many cases, what the user desires is not all of the potentially relevant documents, but, say, only one useful one. The concept of utility as an evaluation measure in a sense is recognition of the importance of taking account of desire on the part of the user. If these user-related factors are ignored, then the evaluation measures which depend upon them, although certainly measuring something, may not be measuring anything practically useful.

Confounded concepts

Satisfaction (of need, of desire) is of course the basic concept in information retrieval system evaluation, and as such cannot be ignored in any test. The various concepts of relevance which have been proposed and used testify to its importance, and to its intractability. There appear to be two strong reasons for making sure that the operational definition of satisfaction is closely related to user judgements. One is, that if user judgement is factored out, then the basis for evaluation of system performance may be unrelated to real situations. The other is that it seems clear that actual satisfaction judgements are order-dependent, and this cannot be dealt with unless one works within a context in which needs are assumed to change with new information. This last point is especially difficult to deal with in any testing environment, whether one recognizes its importance or not, and appears to require the development of some quite new experimental paradigms and evaluation measures.

4.6 Conclusion

This chapter has not tried to provide a list or grab-bag of techniques which could be used to investigate the 'ineffable' concepts of information retrieval. Rather, the aim has been to indicate the ways in which these concepts are important to the information retrieval situation, some reasons for their peculiar intractability, and some reasons for why they need to be considered in the testing of information retrieval systems. Perhaps the discussion and examples have indicated some special problems that these concepts pose: this chapter has been written in the hope that these problems will be taken seriously into account in the design and conduct of future information retrieval system testing.

There are a number of specific conclusions which I think can be drawn from this discussion of ineffable concepts in information retrieval. The first is that studies of these variables *per se* are very much needed, in order to provide a sounder basis for evaluation measures of information retrieval systems, as well as to provide sounder design principles for information retrieval systems. It also seems that evaluation of information retrieval systems must now begin to take more explicit account of the nature of these concepts, especially the user-related group, as problems of evaluation of online systems become more acute. And finally, in order to become explanatory and predictive, rather than merely descriptive, information retrieval system evaluation should change its emphasis (as, indeed, it is already beginning to do). Testing or evaluation has largely been on the basis of describing and comparing the results of system performance, without a great deal of emphasis on the theory underlying the systems. But if information retrieval systems design is to progress in a meaningful sense, then we need theories which allow us to explain *why* one system works better than another. Except in what seem to be relatively minor ways, this cannot be achieved without taking account of at least the variables discussed in this chapter. To that extent, it seems to me that the future of information retrieval system testing and design lies necessarily in the investigation of these concepts and their application.

4.7 Acknowledgement

Much of this work was accomplished while I was on a visiting appointment at the School of Library and Information Science, University of Western Ontario. I would like to thank the School and my colleagues and students there for their help (which they may not recognize).

References

1. BELKIN, N. J. and ROBERTSON, S. E. Information science and the phenomenon of information, *Journal of the American Society for Information Science* **27**, 197–204 (1976)
2. ROBERTSON, S. E. Indexing theory and retrieval effectiveness, *Drexel Library Quarterly* **14**, 40–56 (1979)
3. PAISLEY, W. J. and PARKER, E. B. Information retrieval as a receiver-controlled communication system. In: *Education for Information Science* (Ed. L. P. Heilprin, B. E. Markuson and F. L. Goodman), pp. 23–31, Macmillan, London (1965)

4. BELKIN, N. J. The problems of 'matching' in information retrieval. In: *Theory and Applications of Information Research*, Proceedings of the Second International Research Forum in Information Science, (Ed. O. Harbo and L. Kajberg), Mansell, London (1980)
5. TAYLOR, R. S. Question negotiation and information seeking in libraries, *College and Research Libraries* **29**, 178–89 (1968)
6. WILSON, T. D. and STREATFIELD, D. R. Information needs in local authority social services departments: an interim report on project INISS, *Journal of Documentation* **33**, 277–93 (1977)
7. WERSIG, G. *Information—Kommunikation—Dokumentation*, Verlag Dokumentation, Pullach bei München (1971)
8. BELKIN, N. J. *A Concept of Information for Information Science*, Ph.D. Thesis, University of London (1977)
9. BELKIN, N. J. and ODDY, R. N. Document retrieval based on the automatic determination of the user's information need, *Journal of Informatics* **2**, 8–12 (1978)
10. JAHODA, G. Reference question analysis and search strategy development by man and machine, *Journal of the American Society for Information Science* **24**, 139–44 (1974)
11. LYNCH, M. J. *Reference Interviews in Public Libraries*, Ph.D. Thesis, Rutgers University (1977)
12. BELKIN, N. J. Information concepts for information science, *Journal of Documentation* **34**, 55–85 (1978)
13. FARRADANE, J. The nature of information, *Journal of Information Science* **1**, 13–17 (1979)
14. PRATT, A. D. The information of the image, *Libri* **27**, 204–20 (1977)
15. HUTCHINS, W. J. On the problem of 'aboutness' in information retrieval, *Journal of Informatics* **1**, 17–35 (1977)
16. MARON, M. E. On indexing, retrieval and the meaning of about, *Journal of the American Society for Information Science* **28**, 38–43 (1977)
17. ROBERTSON, S. E. Between aboutness and meaning. In: *The Analysis of Meaning: Informatics 5* (Ed. M. MacCafferty and K. Gray), pp. 202–5, Aslib, London (1979)
18. SPARCK JONES, K. Problems in the representation of meaning in information retrieval. In: *The Analysis of Meaning: Informatics 5* (Ed. M. MacCafferty and K. Gray), pp. 193–201, Aslib, London (1979)
19. HILLMAN, D. J. Customized user services via interactions with LEADERMART, *Information Storage and Retrieval* **9**, 587–96 (1973)
20. GARDIN, J.-C. On the relation between question-answering systems and various theoretical approaches to the analysis of text. In: *The Analysis of Meaning: Informatics 5* (Ed. M. MacCafferty and K. Gray), pp. 206–20, Aslib, London (1979)
21. SARACEVIC, T. Relevance: A review of and a framework for the thinking on the notion in information science, *Journal of the American Society for Information Science* **26**, 321–43 (1975)
22. KEMP, D. A. Relevance, pertinence and information system development, *Information Storage and Retrieval* **10**, 37–47 (1974)
23. WILSON, P. Situational relevance, *Information Storage and Retrieval* **9**, 457–71 (1973)
24. COOPER, W. S. A definition of relevance for information retrieval, *Information Storage and Retrieval* **7**, 19–37 (1971)
25. COOPER, W. S. On selecting a measure of retrieval effectiveness, parts 1 and 2, *Journal of the American Society for Information Science* **24**, 87–100 and 413–24 (1973)
26. KARLGREN, H. Homeosemy: on the linguistics of information retrieval. In: *Natural Language in Information Science* (Ed. D. E. Walker, H. Karlgren and M. Kay), (FID Publication 551), pp. 167–81, Skriptor, Stockholm (1977)
27. ROBERTSON, S. E. *A Theoretical Model of the Retrieval Characteristics of Information Retrieval Systems*, Ph.D. Thesis, University of London (1976)
28. YOVITS, M. C. A theoretical framework for the development of information science. Information science: its scope, objects of research and problems. In: *Problems of Information Science* (FID Publication 530), pp. 90–114, VINITI, Moscow (1975)
29. BELKIN, N. J., BROOKS, H. M. and ODDY, R. N. Representation and classification of anomalous states of knowledge and information for use in interactive information retrieval. In: *IRFIS 3* Proceedings of the Third International Research Forum in Information Science (Ed. T. Henriksen), pp. 146–83, Statens Bibliotekskole, Oslo (1979)
30. BELZER, J. Information theory as a measure of information content, *Journal of the American Society for Information Science* **24**, 300–4 (1973)
31. SALTON, G. (Ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, N.J. (1971)

32. MCGILL, M. J. Knowledge and information spaces: implications for retrieval systems, *Journal of the American Society for Information Science* **27**, 205–210 (1976)
33. COOPER, W. S. The paradoxical role of unexamined documents in the evaluation of retrieval effectiveness, *Information Processing and Management* **12**, 367–76 (1976)
34. ODDY, R. N. Information retrieval through man-machine dialogue, *Journal of Documentation* **33**, 1–14 (1977)