A NEW THEORETICAL FRAMEWORK FOR INFORMATION RETRIEVAL

C J van Rijsbergen
Department of Computing Science
University of Glasgow

ABSTRACT

A new framework based on a non-classical logic is proposed for investigating IR. The paper motivates the use of a particular conditional logic as the 'right' logic for IR. A new principle, the logical uncertainty principle, is proposed, to deal with the inherent uncertainty associated with applicable inferences.

1. PREFACE

In the last few years, I have become increasingly dissatisfied with the state-of-the-art in information retrieval. I have reluctantly concluded that the fundamental basis of all the previous work is wrong. Almost all of the previous work in Information Retrieval (including my own) has been based on the assumption that a formal notion of meaning is not required to solve the information retrieval problems. Typically, researchers have assumed that one could get by, by only considering absence or presence of word tokens in text together with counting information about the distribution of words. Although such an approach has been successful up to a point, it has become clear that further advances in the effectiveness of retrieval by such techniques are not possible. My observation is that performance based on statistical techniques has reached its theoretical limit and any attempts to achieve further improvements are a waste of time. This is not to say that systems based on these techniques are not worth building; on the contrary, they are because it is the best we have to date. But to build a new generation of Information Retrieval systems, a new theory will be needed.

2. INTRODUCTION

This paper is to be seen as describing a new theoretical framework for investigating information retrieval. The details at this stage are not completely worked out; some further details are given in Van Rijsbergen (1986). For some years now, I have felt the need to describe such a framework. It is especially important if one wants to develop information retrieval beyond the mere keyword approach. In the closing pages of my earlier book on Information Retrieval (Van Rijsbergen, 1979) I said the following: "It has never been assumed that a retrieval system should attempt to 'understand' the content of a document. Most Information Retrieval systems at the moment merely aim at a bibliographic search.

Documents are deemed to be relevant on the basis of a superficial description. I do not suggest that it is going to be a simple matter to program a computer to understand documents. What is suggested is that some attempt should be made to construct something like a naive model, using more than just keywords, of the content of each document in the system. The more sophisticated question-answering systems do something very similar. They have a model of their universe of discourse and can answer questions about it, and , can incorporate new facts and rules as they become available".

When I wrote the above passage, I had no idea that progress in that direction was going to be so slow. The main obstacles appeared to be an adequate computable model of meaning, and its use in information retrieval operations. It was argued that even if we had an appropriate semantics for text, and it could be computed efficiently, we still would not know how to use it to retrieve documents in response to requests.

I would now like to counter this objection by saying that the use of semantics comes via an appropriate logic. I am not alone in thinking this; Cooper (1978) in his book on logico-linguistics would probably make the same claim. Such a logic would be based on a formal semantics for text. The semantics would provide a limited representation of the meaning of any text but it would not be the meaning. A logic would then be interpretable in that semantics. It leaves me to say how such a logic can help in the retrieval of relevant documents. To understand this, one must think of documents as sets of sentences which are interpreted in the semantics, and think of queries as sentences too, the latter usually a single sentence. The single primitive operation to aid retrieval is then one of uncertain implication. In the extreme case, it would be logical implication, which through its interpretation in the formal semantics is logical consequence. That is, a document is retrieved if it logically implies the request. However, as we all know, documents rarely imply requests; there is always a measure of uncertainty associated with such an implication. And so, a notion of probable, or approximate, implication is needed where a plausible inference instead of a strict inference is made, and the plausibility quantified through some measure. Modelling the information retrieval process in this way goes beyond the keyword approach, and specifies, once and for all, what relationship between a document and a request is to hold to compute probable relevance. The importance of this new way of looking at Information Retrieval derives from the realisation that with such a framework, Information Retrieval can advance with new developments in formal semantics for text. Starting with a keyword analysis which is a primitive semantics, we can go on to use our logic no matter how sophisticated our semantics is. At all times, we are attempting to infer requests (treated as sentences) from statements in the documents. The inference is possible because we have an interpretation of sentences in a document, we define this interpretation and can increase its complexity at will.

It is important to realise that the above approach is similar to the one adopted in database querying and question-answering. It is similar in that in all cases, the answer is obtained through a process of logical satisfaction, i.e. looking at a common interpretation for premises and consequent. It is different in that in the case of Information Retrieval, a request is typically a closed sentence (i.e. contains no variables) and the relationship computed between a document (the premises) and the request (the consequent) is paramount i.e. if the relationship is sufficiently strong, the document is retrieved. In the case of Data Base Management Systems, a request is typically an open sentence (contains variables), the semantics giving an instantiation of the request, which is an answer.

3. CLASSICAL INFORMATION RETRIEVAL

To begin with, I would like to say what Information Retrieval is. Let us assume that there is a large store of documents on a variety of topics. A user of such a store will have a need to know certain things, things that he does not know at present. He therefore expresses his information need in the form of a request for information. Information Retrieval is concerned with retrieving those documents that are likely to be relevant to his information need as expressed by his request. It is likely that such a

retrieval process will be iterated, since a request is only an imperfect expression of an information need, and the documents retrieved at one point may help in improving the request used in the next iteration. It is important to realise that certain words in the above description are used carefully to avoid misunderstanding the idea of information retrieval.

Let us spell out the way in which the description is to be interpreted. A request for information is translated into a request for documents. The documents are assumed to contain the information, therefore the information is only retrieved indirectly. A request is an imperfect expression of a user's information need; only a user will be able to tell whether a document contains the information he is seeking. If it does contain the information sought then the document is considered relevant to the user's information need. This implies that documents are not relevant to a request; that is, identical requests submitted by two different users can be satisfied in different ways, one document may be relevant to one user and not to the other. Relevance is here connected firmly to aboutness, a document is not relevant because of its colour or shape. It is relevant because it is about the information sought.

In specifying a model for information retrieval, a small number of entities and concepts need to be defined. Superficially, this would appear to be a simple matter. The entities and concepts are document, request, property of a document, and relevance. Anyone can give commonsense definitions of these; unfortunately, what is required is a formal definition so that an Information Retrieval system can be formally specified and therefore implemented on a computer.

Let us take a document as a set of sentences. Therefore, when a document is considered for retrieval, the sentences in the document are considered individually or perhaps jointly. In considering them, one is looking for a relationship between them and the request. Such a relationship needs to be computable if the Information Retrieval system is a computer-based one. If we take a request to be a sentence then the relationship to be computed is one between a set of sentences and a single sentence. This relationship must be such that it enables one to use it to determine whether a document is likely to be relevant or not. I use 'likely' because we are assuming that relevance is user dependent and a request is an imperfect expression of an information need.

From a system's point-of-view, the computation of the relationship between document and request is central. How is one to specify this relationship? There are several ways of doing this, each one has implications for how one represents a document and a query. Ideally, one would like this representation to be separated from the relationship computation; of course, this has proved to be almost impossible. In what follows, I propose that the right representation is given by a formal semantics for text (perhaps a Montague-style semantics, see Dowty, et al, 1981). The detailed specification of a semantics will be the subject of a later paper. The relationship between a document and a request will be formalised as a logical implication to which a measure of uncertainty is attached. To motivate this 'implication' I shall give two examples in which standard Information Retrieval models are re-expressed in terms of uncertain implication.

4. BOOLEAN RETRIEVAL

It is assumed that documents are represented by index terms, or keywords, and that requests are logical combinations (using AND, OR, NOT) of these terms. A document is deemed likely to be relevant, and hence retrieved, if the index terms in the document satisfy the logical expression in the request. For example:

D_1 = [A,B]

D_2 = [B,C] A,B,C : index terms

D_3 = [A,B,C]

$Q = A \wedge B \wedge C$

D_1 : retrieved because D_1 is true implies Q is true

D_2, D_3 : not retrieved.

The index terms are, in fact, the semantics, and indexing is seen as mapping a piece of text into its formal semantics. Formally, an index term is true for a document if it occurs in the set representing the document.

Notice the use of the closed world assumption here, that the absence of an index term in a document is assumed to imply that it is false for that document. The example makes clear that the relation computed between D and Q is one of logical implication. This is a simple set-up and commonly used in practice. Unfortunately, it does not model the uncertainty of relevance.

5. CO-ORDINATION LEVEL MATCHING

Just as in the example of Boolean retrieval above, documents are assumed to consist of sets of index terms, but requests are now also sets of index terms. The relationship between a document and a request is now computed in terms of the index terms they have in common. The likelihood of relevance is taken to be directly proportional to the number of index terms shared. For example, D_1, D_2, D_3 as before,

$Q = [A,B,C]$:

$n(D_1 \cap Q) = 2$

$n(D_2 \cap Q) = 2$ $\qquad$ $n(.)$: number in set

$n(D_3 \cap Q) = 3$

This relationship can be described in terms of the probability of a logical implication, so that, $n(D \cap Q)$ is proportional to the probability of $D \rightarrow Q$. What is a probability of $D \longrightarrow Q$? This firstly depends on how one interprets '$\longrightarrow$'. It is not to be interpreted as the material implication: $D \supset Q$, which is the usual truth-functional connective, only false when D is true and Q is false. Intuitively, whatever the precise meaning of '$\longrightarrow$', it is easy to understand that $D \longrightarrow Q$, or that $D \not\longrightarrow Q$. The problem is that when $D \not\longrightarrow Q$ we might still want to retrieve D because of its likelihood of relevance. To model this uncertainty of relevance, we use uncertainty of implication. If we assume $P(D \longrightarrow Q) = P(Q \mid D)$, then with D and Q as sets we have:

$$P(D \longrightarrow Q) = \frac{P(Q \cap D)}{P(D)} = \frac{n(Q \cap D)}{n(D)}$$

Treating $n(D)$ as constant, we get the relationship that $P(D \longrightarrow Q)$ is proportional to the level of co-ordination.

6. A CONDITIONAL LOGIC FOR INFORMATION RETRIEVAL

In re-expressing the two well known retrieval models, Boolean and Co-ordination, as examples of computation of logical implication, I have made the case (in part) that the fundamental retrieval operation is one of logical implication. Probabilistic retrieval can also be modelled as uncertain logical implication. For this the reader is referred to Van Rijsbergen (1986). This logical implication is not one of material implication, the usual truth-functional connective $A \supset B$, which is true in all cases except when A is true and B is false. To illustrate the difference between our earlier implication $A \longrightarrow B$ and $A \supset B$

let me give a simple example. First, let us assume that the probability of a conditional of the form 'If A is true then B ' is a conditional probability. Now consider a die and two events, A the event 'a number less than 3 will be rolled' and B the event 'an even number will be rolled'.

Then for the two 'implications' we get:

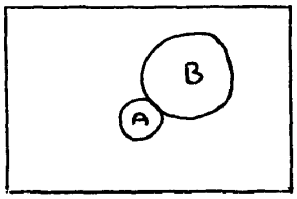
$$P(A \rightarrow B) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{2/6} = \frac{1}{2}$$

$$P(A \supset B) = P(\bar{A} \cup B) = \frac{5}{6}$$

This shows that by interpreting the probability of a conditional as a conditional probability rather than the probability of a material implication we get widely differing results. Of course, I would maintain that the conditional probability interpretation in the context of Information Retrieval is the right one.

There is another major reason why a conditional must not be identified with the material implication in logic. When using probabilistic inference, we want to ensure that the following soundness criterion holds (Adams, 1975) : It is impossible for the premises of an inference to be probable while its conclusion is improbable. To illustrate a violation of this, we take the well known inference: Given $\sim A$ we can infer $A \supset B$. [Remember that we can logically infer a consequent from an antecedent, whenever interpretations making the antecedent true also make the consequent true.] In our example, whenever $\sim A$ is true, A will be false and hence $A \supset B$ will be true independent of B's truth value.

If we identified $A \rightarrow B$ with $A \supset B$, then such an inference could easily violate the soundness criterion. It is easy to show situations (see diagram below) where $P(\sim A)$ is large and $P(A \rightarrow B) = P(B \mid A)$ (probability of consequent) is small. In other words, although '$\sim A$ infer $A \supset B$ ' is valid '$\sim A$ infer $A \rightarrow B$ ' should not be, if we take the probabilistic soundness criterion seriously.



$P(\sim A)$ large
$P(B \mid A) \rightarrow 0$

A conditional logic will, therefore, in general, be different from a classical logic (Harper, et al, 1981). It is my contention that such a conditional logic (and there are several formulations) is the correct one for information retrieval.

7. HOW DO WE EVALUATE $P(s \rightarrow q)$?

First, let us consider the case without probabilities. To analyse this case, we will need to introduce possible-world semantics. An intuitive understanding of a possible world is that it is a complete specification of how things are, or might be, down to the finest semantically relevant details (Bradley and Schwartz, 1979). For our purposes, we will identify documents with possible worlds. This will raise problems of finiteness and structure which we will ignore for the moment.

Let s be a partial description of a document, this might be a set of sentences, or just a single index term, q will be a request. In deciding whether to retrieve a document, we would need to evaluate $s \rightarrow q$,

that is, whether $s \rightarrow q$ is true or not. If s is true in a document d then $s \rightarrow q$ is true providing q is true. If s is not true in a document then we go to the nearest document d' to d in which it is true and consider whether q is true. If q is true in d' then $s \rightarrow q$ is true in d, otherwise it is false.

To give a simple example, s might be an index term, q the same or a different index term. If $s = q$, $s \rightarrow q$ is true follows trivially for those documents in which q occurs. The more interesting case is when $s \neq q$. In this case, to establish $s \rightarrow q$ in d find the nearest document d' in which s occurs and check for the occurrence of q. It is important to realise that because of the primitive nature of the semantics, an example such as s = FORTRAN, q = PROGRAMMING LANGUAGE for which $s \rightarrow q$ is directly true in a more complex semantics, can only be handled indirectly.

The above process illustrates what is now widely known as the Ramsey test (Mellor, 1976). It might be summarised as follows:

> To evaluate a conditional, first hypothetically make the minimal revision of your stock of beliefs required to assume the antecedent. Then, evaluate the acceptability of the consequent on the basis of this revised body of beliefs.

Note the meaning of a conditional is not truth-functional under the above interpretation, i.e. its truth does not simply depend on the truth valuation of s and q in one world. It has become an intensional notion.

In document retrieval, we are often faced with the situation where $s \rightarrow q$ is assumed false because s does not logically imply q. That is, assuming the truth of the sentences (index terms) in a document we cannot arrive at q. Boolean retrieval is an excellent example: given a truth valuation for the terms describing a document, we retrieve those documents which imply q (make q true for that valuation). What is suggested here, is that a given document should be revised in a minimal way that makes s true. If, after that revision, q is true, then $s \rightarrow q$ is true and d should be retrieved. There are a number of ways of making this revision. One could restrict the revision to selecting a nearest document in which s is true, in which case, no interaction from the user would be required. Or, one could involve the user in expanding the information contained in the document under consideration. Or, finally, one could do document expansion automatically using information already stored in the system. We will return to this notion of minimal revision when we attempt to formalise it.

Turning now to the probabilistic case, to evaluate $P(s \rightarrow q)$, we revise the probability function P to P' in a minimal way, so that $P'(s) = 1$. We then have that:

$$P(s \rightarrow q) = P'(q) .$$

An example of such a revision is to make $P(s \rightarrow q) = P(q \mid s)$. In the case of Boolean semantics, where x,y are index terms and v a truth valuation:

$$v(x) = \begin{cases} 0 \\ 1 \end{cases} \qquad\qquad v(y) = \begin{cases} 0 \\ 1 \end{cases}$$

we get

$$P(x \rightarrow x) = 1$$
$$P(y \rightarrow x) = P(x \; y)$$

In other words, a query consisting of the index term x, is related to a document containing y by $P(x \mid y)$. If we restrict our worlds to documents already present, then we can interpret this as:

$$\frac{n(x \wedge y)}{n(y)}$$

the frequency of the co-occurence of x and y divided by the frequency of y.

Of course, documents and queries are far more complex than is assumed above. It is not clear yet how one deals with arbitrary complex documents and queries. Generalising from the simple index term approach we would need to specify a formal semantics in which documents and queries would be interpreted. To evaluate $s \longrightarrow q$ would require a change in the interpretation function so that s would be true under the new interpretation, and $s \not\longrightarrow q$ true, if q was true as well.

8. LOGIC OF UNCERTAINTY

In evaluating the truth of $y \longrightarrow x$ or evaluating $P(y \longrightarrow x)$, we are dependent on a notion of nearness (closeness) between worlds or documents. It is interesting to examine this in a little more detail. Remember our prime concern is to establish that '$y \longrightarrow x$', or that $y \longrightarrow x$, with sufficiently large probability. If for the current document $y \not\longrightarrow x$, we look at the effect of changing/revising our current world and look at $y \longrightarrow x$ in the revised world. These changes are to be made in a minimal way.

There is another way of looking at this revision process which may be more appropriate in the Information Retrieval context. I would like to generalise the Ramsey test and state a new principle.

Logical Uncertainty Principle

> Given any two sentences x and y; a measure of the uncertainty of $y \longrightarrow x$ relative to a given data set, is determined by the minimal extent to which we have to add information to the data set, to establish the truth of $y \longrightarrow x$.

This is a slight generalisation of the foregoing. It denies that one can assess $y \longrightarrow x$ with certainty if one has to revise the data set. It says nothing about how 'uncertainty' or 'minimal' might be quantified. It specifically relativises truth to a given data set. The semantics of the data have been left unspecified too. Nearness has been replaced by a measure of information. How this measure might be evaluated is suggested in Van Rijsbergen (1986).

9. CONCLUSION

In this paper, I have given a new framework for Information Retrieval based on non-standard logic. The fundamental primitive operation relating documents and queries is taken to be logical implication. This is not a truth functional notion in the classical sense, but rather can only be evaluated by considering truth in other possible worlds. A new logical uncertainty principle is stated to characterise uncertainty associated with any logical implication; thereby quantifying the uncertainty of relevance.

ACKNOWLEDGEMENTS

I would like to thank Bruce Croft for his helpful comments in writing this paper.

REFERENCES

Adams, E.W. (1975) The logic of conditionals, Dordrecht: Reidel.
Bradley, R. and Schwartz, N. (1979) Possible Worlds, Oxford: Basil Blackwell.
Cooper, W.S. (1978) Foundations of Logico-Linguistics, Dordrecht: Reidel.
Dowty, D.R., Wall, R.E. and Peters, S. (1981) Introduction to Montague Semantics, Dordrecht: Reidel.
Harper, W.L., Stalnaker, R. and Pearce, C. (1981) (eds) Ifs, Dordrecht: Reidel.
Mellor, D.H. (1976) (ed) Foundations: Essays in Philosophy, Logic, Mathematics and Economics: F P Ramsey, London: R.K.P.
Van Rijsbergen, C.J. (1979) Information Retrieval, 2nd edition, London: Butterworths.
Van Rijsbergen, C.J. (1986) A non-classical logic for information retrieval, The Computer Journal (in press).