A non-classical logic for information retrieval

C. J. VAN RIJSBERGEN*
Department of Computer Science, University College, Dublin 4

Implicit in many information retrieval models is a logic. These logics are hardly ever formalised. This paper formalises a non-classical logic underlying information retrieval. It shows how a particular conditional logic is the 'right' logic to do Information Retrieval. Its relationship to existing retrieval mechanisms is investigated. The semantics of the logic are expressed in probability theory, and evaluated through a possible-world analysis, thus establishing an intensional logic. In doing so, we motivate a new principle, the* logical uncertainty principle, *which gives a measure of the uncertainty associated with an inference.

Received January 1986

1. INTRODUCTION

This paper is to be seen as describing a new theoretical framework for investigating information retrieval. 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 the subject 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 naïve 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.'[1]

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, in his book on logico-linguistics, would probably make the same claim.[2] 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 variable) 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.

Editor's Note: It may be thought that the publication of this paper is premature, in that the presentation of the proposed logic is not supported by rigorous mathematical argument. The purpose of publishing now is, however, to provoke reaction to a completely new direction for research in Information Retrieval. Responses will be especially welcome.

* Address for correspondence: Computing Science Department, The University, Glasgow G12 8QQ.

2. 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 twó 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, and 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 Ref. 3). 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 three examples in which standard Information Retrieval models are re-expressed in terms of uncertain implication.

2.1. 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
$D_3 = \{A, B, C\}$
$Q = A \wedge B \wedge \sim C$
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.

2.2. 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,

$$Q = \{A, B, C\}:$$

$$n(D_1 \cap Q) = 2$$
$$n(D_2 \cap Q) = 2$$
$$n(D_3 \cap Q) = 3$$

where $n(\;)$ = number in set.

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 $D \rightarrow Q$? This depends first on how one interprets '$\rightarrow$'. 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 '→', it is easy to understand that $D \rightarrow Q$, or that $D \nrightarrow Q$.

The problem is that when $D \nrightarrow Q$ 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 $P(D \rightarrow Q) = P(Q \mid D)$, then with D and Q as sets we have:

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

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

2.3. Probabilistic Retrieval

In this example, documents are also represented by sets of index terms, and so are queries. However, this time the relationship between them is calculated by including estimates of the likelihood that a shared term indicates relevance. The emphasis is on somehow finding out how index terms discriminate between relevant and non-relevant documents. For example, a user might indicate that an index term is a good discriminator, i.e. it occurs far more frequently in relevant than in non-relevant documents. Such information for a number of terms is then pooled to estimate the probability of relevance of a particular document.

Consider a document represented by D that has not been retrieved before, its probability of relevance being given by $P(\text{rel} \mid D)$. This probability is assumed to be well formed in the sense that 'rel' and 'D' are events or propositions for which the relationship of probability holds. Unfortunately, this is not so; 'rel' is neither a proposition nor an event. Relevance is only given after the event of retrieval, and is a function of the user. Therefore, relevance can be used to conditionalise probabilities, but it cannot be given equal status with documents and requests, which are known before a retrieval operation.

Now, although 'D' appears as a simple event in $P(\text{rel} \mid D)$ its interpretation is far from simple. In the standard probability model we assume D to be a vector-valued random variable,[1] where its distribution is given by a mixture of two distributions, namely

$$P(D) = P(D \mid \text{rel})\, P(\text{rel}) + P(D \mid \text{nrel})\, P(\text{nrel})$$
($n$rel = not-relevant).

To compute $P(\text{rel} \mid D)$ we use Bayes' Theorem:

$$P(\text{rel} \mid D) = \frac{P(D \mid \text{rel})\, P(\text{rel})}{P(D)}$$

In this computation the relationship between a document description and a request is given only indirectly. The request is used to start the iterative process in evaluating $P(\text{rel} \mid D)$. On the first cycle, one needs an estimate of $P(D \mid \text{rel})$, which can be obtained by using the request to retrieve some documents and assessing them for relevance. Another way of putting this is that P is *revised* to a different probability function P_{rel} in the light of information about relevance, and that

$$P(D \mid \text{rel}) = P_{\text{rel}}(D).$$

Putting it this way makes it clear that two users with differing ideas of relevance but submitting the same request can expect to get different probabilities of relevance, i.e. user 1 would get $P^1_{\text{rel}}(D)$ and user 2 $P^2_{\text{rel}}(D)$. This simply means that the probability function P can be revised in two different ways. But what about the case of the same relevance judgements but different requests, e.g. q_1 and q_2? As it stands, the probabilistic model does not deal with it directly. A recent attempt to deal with it can be found in (Ref. 4).

I would like to propose the following way of dealing with both cases; different relevance judgements and different requests.

Instead of calculating $P(D)$ or $P_{\text{rel}}(D)$, I propose $P(s \rightarrow q)$ or $P_{\text{rel}}(s \rightarrow q)$. Here s is a description of a document (for example, a set of sentences) and q a description of a request. $s \rightarrow q$ is a logical implication and $P(s \rightarrow q)$ is a measure of its uncertainty. In doing this, we have done two things: (1) separated the process of revising probabilities from the logic; and (2) separated the treatment of relevance from the treatment of documents and requests.

The general picture we now have is that the probability of relevance is given by the probability that q follows from s. However, this latter probability is a function of what the user already knows. His knowledge is expressed through relevance judgements and quantified through revision of the P to P_{rel}.

3. A CONDITIONAL LOGIC FOR INFORMATION RETRIEVAL

In re-expressing the three well-known retrieval models, Boolean, Co-ordination and Probabilistic, as examples of computation of logical implication, I have made the case (in part) that the fundamental retrieval operation is one of logical implication. 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 $A \rightarrow B$ 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:

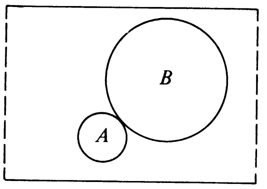
$$P(A \rightarrow B) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{2/6} = \frac{1}{2}$$

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

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.[5] 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.[6] It is my contention that such a conditional logic (and there are several formulations) is the correct one for information retrieval.

4. 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.[7] 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 being 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* .[8] 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 that 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:

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

we get

$$P(x \rightarrow x) = 1$$
$$P(y \rightarrow x) = P(x \mid y).$$

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-occurrence 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 $s \rightarrow q$ would require a change in the interpretation function so that s would be true under the new interpretation, and $s \rightarrow q$ true, if q was true as well.

5. LOGIC OF UNCERTAINTY

In evaluating the truth of $y \rightarrow x$ or evaluating $P(y \rightarrow x)$, 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 '$y \rightarrow x$', or that $y \rightarrow x$, with sufficiently large probability. If for the current document $y \nrightarrow x$, we look at the effect of changing/revising our current world and look at $y \to x$ 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, Logical Uncertainty Principle.

> Given any two sentences x and y; a measure of the uncertainty of $y \to x$ 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 $y \to x$.

This is a slight generalisation of the foregoing. It denies that one can assess $y \to x$ 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.

Conventionally, uncertainty has been measured in information-theoretic terms. I will do the same. If we restrict ourselves to documents, and identify 'data set' with 'document', then we require an information measure to make the above principle precise. Formulating this, given any two documents w_1, w_2 we define conditional information measures $I(w_1 | w_2)$ and $I(w_2 | w_1)$, which give the information contained in w_2 about w_1 and vice versa. Notice that $I(.|.)$ is not symmetric, although one could define a symmetric *mutual* information:

$$I(w_1 : w_2) = I(w_1) - I(w_1 | w_2) = I(w_2) - I(w_2 | w_1).$$

The details are not important. What *is* important is that $I(.|.)$ can be used as a nearness measure, and that it can be defined *algorithmically* without recourse to random variables.[9] How is this done? Essentially, the conditional information measure $I(w_1 | w_2)$ is defined to be the smallest program needed to calculate w_1 from a minimal program for w_2. Now we have a nearness measure in terms of the information contained in one object about another. Given a document w, to find the nearest document in which a sentence is true we find that α for which $I(\alpha | w)$ is a minimum, subject to the sentence being true in α. In an intuitive sense this is the least revision of the given document, i.e. requires the smallest program to calculate α.

Of course, the principle does not specify that further information should come from the document collection. It may be that a thesaurus, or an expert assistant will be the source of the extra information. The revised document will then probably be different from any document already present. Let us consider an example of the second kind, an expert assistant.[10] Such an assistant might contain rules such as:

if a **then** b [0.9] i.e. $P(a \to b) = 0.9$.

If the query is y and the document contains x, then to derive the probability of $x \to y$, we would have to find intermediate steps. For example:

$$P(x \to a), \quad P(a \to b), \quad P(b \to y).$$

Each of these steps is either given by an expert assistant or can be evaluated from the document collection. How one combines these separate pieces of evidence to give a value for $P(x \to y)$ remains an open question. The reason it is an open question is related to the problematic status of $x \to y$ as a logical proposition and the consequent impossibility of simply embedding propositions of this kind. It is not clear that $a \to (b \to d)$ can be treated as $P(b \to d | a)$.[11] Clearly, one would like to do this, but simple approaches have led to the identification of $b \to d$ with $b \supset d$. At this stage I would *conjecture* that if one used a probability revision proposed by Lewis,[6] embedding would be allowed and would not lead to paradoxical results. However, this revision process may not be acceptable on other grounds. It would appear to me that a specification of a formal semantics for '$\to$' would be the way forward; it is the subject of a paper in preparation.

6. 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.

REFERENCES

1. C. J. van Rijsbergen, *Information Retrieval*, 2nd edition. Butterworths, London (1979).
2. W. S. Cooper, *Foundations of Logico-Linguistics*. Reidel, Dordrecht (1978).
3. D. R. Dowty, R. E. Wall and S. Peters, *Introduction to Montague Semantics*. Reidel, Dordrecht (1981).
4. S. E. Robertson, M. E. Maron and W. S. Cooper, Probability of relevance: a unification of two competing models of document retrieval. *Information Technology: Research and Development* **1**, 1–21 (1982).
5. E. W. Adams, *The Logic of Conditionals*. Reidel, Dordrecht (1975).
6. W. L. Harper, R. Stalnaker and C. Pearce (eds), *Ifs*. Reidel, Dordrecht (1981).
7. R. Bradley and N. Schwartz, *Possible Worlds*. Basil Blackwell, Oxford (1979).
8. D. H. Mellor (Ed.) *Foundations: Essays in Philosophy, Logic, Mathematics and Economics: F. P. Ramsey*. Routledge & Keegan Paul, London (1976).
9. G. J. Chaitin, Algorithmic Information Theory. *IBM Journal of Research and Development* **21**, 350–359, 496 (1977).
10. W. B. Croft, *User-specified Domain Knowledge for Document Retrieval* 1986 (in press).
11. A. Appiah, Generalising the probabilistic semantics of conditionals. *Journal of Philosophical Logic* **13**, 351–372 (1984).