INFORMATION RETRIEVAL AND INFORMATIVE REASONING

C J VAN RIJSBERGEN

Rapporteur: Cecilia Calsavara

Information Retrieval and Informative Reasoning

C.J. van Rijsbergen
Glasgow University
U.K.

Introduction

Information Retrieval has been a bit of a Cinderella subject, claimed by the Library and Information Science community and viewed with a certain amount of suspicion by the Computer Science community (why not solve it by DB technology?[1]), but as part of computing its roots go back to at least before World War II when Robert Fairthorne was speculating about the use of computing machinery to enhance the retrieval of bibliographic records. The physical storage of large amounts of information on electronic media has ceased to be a major problem, hence the emphasis in IR on the *retrieval* of the stored information. The arrival of electronic storage devices that will comfortably handle data sets in the terrabyte range has allowed researchers and developers to concentrate on the searching, retrieval, browsing and display of multi-media information.

What is the Information Retrieval problem? Why is it not enough to say: 'Just store it and when you need it just find it, and retrieve it!' In fact as long ago as in Plato's day the IR problem was already apparent. I quote (paradoxically):

> 'And how will you enquire, Socrates, into that which you do not know?...if you find what you want, how will you ever know that this is the thing which you did not know?...a man cannot enquire either about that which he knows, or about that which he does not know: for if he knows, he has no need to enquire; and if not, he cannot; for he does not know the very subject about which he is to enquire.' Meno, Plato

The storage step is now easy. Finding the right information and looking at it is a different matter. A user interested in a item of information, first has to ask for it. In IR asking for it means constructing a query which may be a natural language statement, a tune, a picture, a bit of image, etc. This query is formulated to reflect or represent what the user is interested in, and ultimately intended to lead to items that the user

[1] But see Harper and Walker's work

wishes to explore, read, think about, absorb etc. The process is one of locating information for further examination.

The Haystack Analogy

In practice the number of items[2] in the repository of items likely to be of interest to a user is small, hence one is dealing with the computation of very rare events. It is like looking for a needle in a haystack, this simple analogy to a haystick will give you some idea of the early models that were proposed for solving the IR problem. If you were given the problem of finding a needle in a haystack you would very quickly find a way to do it. To begin with you would think about the problem in the following way:

1. the thing I am looking for is rare, so a random search is likely to be useless
2. its characteristics are very different from the things I do not want (hay)
3. I can exploit this difference if I know more about it, I could use a magnet or burn down the stack.

Now imagine that the same problem was faced by a Martian who knows nothing about the properties of metal, for example that it is resistant to fire and attracted by magnets. How would such a creature proceed? Well, the Martian's situation is not very different from a user's faced with a large information store: she knows that there is likely to be an item of interest but it is one amongst millions which are of no interest. To take a random sample does not help much for in the case of the haystack it would produce only hay, from which we infer that the needle is not made of straw! We need to find out more about the item looked for, a Martian would consult our knowledge about needles. In the case of IR the user would express a query which when processed would return items likely to be like the ones sought and it is precisely the charateristics of the items returned that could be used to continue the search. The approximate response of the IR system will give clues about the nature of the items sought. It is important to emphasise that what is retrieved is not a random sample.

The QM Analogy

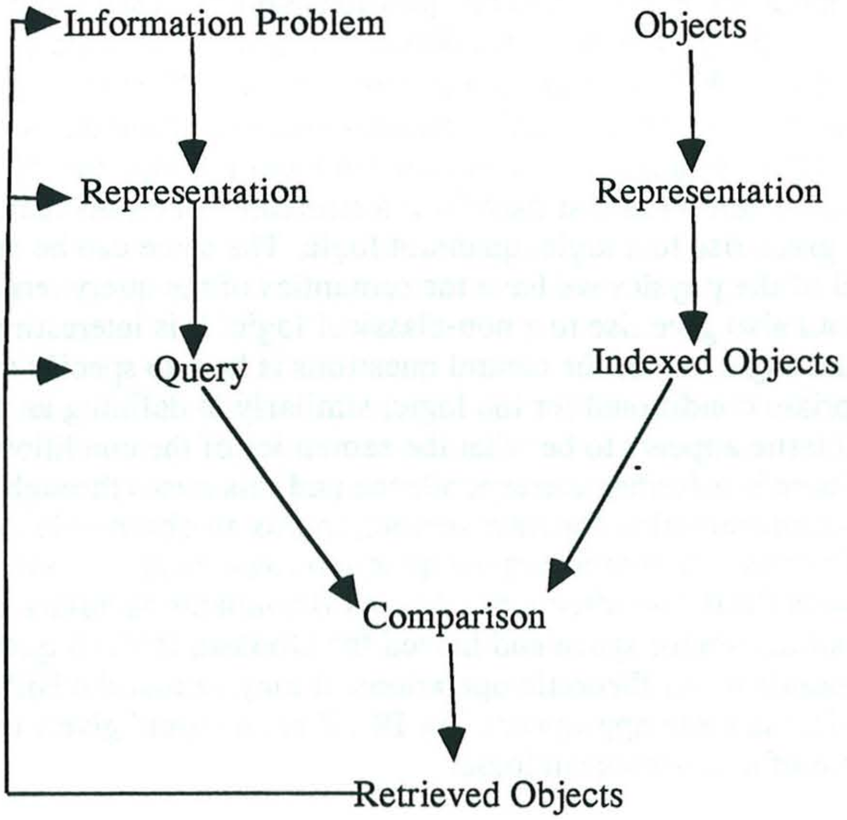
The haystack analogy is a very simple one, and indeed does correspond to some retrieval models. Another analogy which is far more sophisticated and leads into more recent IR models, is one based on the

[2] object=item=document

quantum mechanical paradigm. As many of you will know the theory of quantum mechanics is mainly concerned with observables and state vectors. It is possible to reduce the QM analysis to one in which the observables are simple 'yes'/'no' questions about the state of a system. Moreover the answer to such a binary question has associated with it a probability of being either yes or no in any state. If you now think of the states as documents in an information space and observables as simple queries then the analogy is complete. In fact the properties of simple observables are such that they form a structure, a non-boolean lattice, which gives rise to a logic: quantum logic. The same can be said for IR, instead of the physics we have the semantics of the query terms, and these can also give rise to a non-classical logic. It is interesting that in quantum logic one of the central questions is how to specify an appropriate conditional for the logic; similarly in defining an IR logic the central issue appears to be what the semantics of the conditional will look like. There is a further correspondence and this arises through the Hilbert space respresentation for state vectors, in this an observable can be thought of as a subspace and the quantum logic as operations on subspaces. In IR we often represent our documents as points in a high dimensional vector space and indeed the Boolean logic of queries corresponds to set-theoretic operations, it may be that the Hilbert space formalism is more appropriate for IR , if so, it would give a concrete example of a non-boolean logic.

Feedback and Iteration

How does the IR problem differ from the DB problem? In databases it is asumed that the user can specify completely and accurately the data items of interest. In the relational technology this means asserting a logical combination of attribute values to be satisfied. Any item not satisfying the query in this way is assumed to be of no interest. In IR such an approach simply will not work. Given a logical combination of keywords the chances are that no item will satify it, or that too many will! In either case the answer (the set of items) is not the end of the story. A null answer does not mean that there are no items of interest, nor does a large set as response mean that all the retrieved items are of interest. How does IR get around this problem? The main solution rests on an iterative approach to retrieval, using feedback to focus the search. Conceptually one attempts to locate the *relevant* items in the store proceeds to iteratively discover the attributes of such relevant items so that they can be retrieved.



In statistical terms looking for a rare item is not easy, the literature on signal detection has established that. In terms of the haystack analogy, starting a search might consist of taking a random sample, since this is the time honoured way of estimating properties subject to randomness, but in doing so you are likely to find nothing. To start the iteration one needs to use clues which are likely to tell you something about the items sought. Once such items have been identified their properties can be used to improve the search In other words once a relevant item has been found (because the user says so) one assumes that other relevant items are like it (support fror this comes from the Cluster Hypothesis). Several tools are available to enhance both the intitial and subsequent searches,

- document clustering[3]
- query expansion[4]
- dictionaries and thesauri
- word sense disambiguation
- relevance feedback[5]

[3] See my other paper in these proceedings.

[4] ditto

[5] See later this paper

Retrieval as Inference

In the last few years a new way of looking at the IR problem has been explored: retrieval is modelled as a form of inference. It is simplest to explain this in terms of textual objects. A document is seen as a set of assertion or propositions and a query is seen as a single assertion or proposition[6] Then, a document is considered relevant to a query if it implies the query. The intuition here is that when say q is implied by Δ then Δ is assumed to be about q. Another reason for being attracted to this view of things is that it captures a notion of information containment. When A => B[7] then the information that B is contained in A e.g. A = 'is a square' contains B = 'is a rectangle'. Hence by seeking that which *implies* the query we are seeking that which is *about* the query and that which contains the information specified by the query.

$$\begin{array}{l} If \quad \Lambda \Rightarrow q \quad then \\ \Delta \quad is\,about \quad q \end{array}$$

Although I have represented retrieval as logical inference it is not enough. Basing retrieval on strict logical consequence has the major disadvantage that typically a query is not implied by any documents (of course, this is similar to the failing of Boolean retrieval). Nevertheless it is possible to extend and modify the implication retaining the idea of inferring the query but making it less strict. For this we move to the idea of *partial entailment*, *degree of provability*, or *plausible inference* as it is variously called.

There is an extensive literature on partial entailment going back to at least Leibniz. The intuition is that we are able to assess the degree to which a proposition is entailed by a set of other propositions. So if we have a set Δ , then Δ —> q is measurable. One of the ways is through calculating the conditional probability P(q|Δ). Another way is to evaluate Δ —> q as a conditional whereby we measure the extent to which Δ has

[6] In this paper I will not distinguish between assertions and propositions.

[7] At this point nothing is said about the nature of the implication.

to be augmented so that Δ —> q will go through[8]. The details of this latter approach can be found in my earlier work[9] and that of others[10].

The above approach has much in sympathy with evidential reasoning, that is, given that q constitutes the evidence/clues for the identity of a relevant document D, then we can interpret P(q|D) or P(D|q) as the strenght of evidential support.

Vector Space and Probabilistic Models

Possibly the most complete and satisfactory models for IR that have been described in the literature over the last 20 to 30 years are the vector space and probabilistic models for retrieval.These models are very abstract and assume a considerable data reduction of the objects in the domain of application before they can be used. The vector space model assumes that documents and queries can be represented as points in a high dimensional vector-space and that relevance is measured by a query's proximity to documents in that space. The probabilistic model assumes that one can estimate P(relevance|document). This is done by assuming that the document is a random vector and that relevance is a property whose probability for any unknown document can be estimated by using sample information from a set of known documents.[11] The model then can be shown to be optimal with respect to quality of retrieval if the documents are then retrieved in their order of probability of relevance.[12] Under both models most of the semantics associated with the documents and queries is lost, in fact, it is replaced by absence/presence index terms or frequency counts of those terms.

The basis for Probabilistic Retrieval is the celebrated Bayes' Theorem:

$$P(H|e) \propto P(e|H)P(H)$$

$$\text{or} \quad P(H|e) = \frac{P(e|H)P(H)}{P(e)}$$

$$\text{since} \quad \sum_{H} P(H|e) = 1 .$$

[8] See my other paper in these proceedings.

[9] Logic papers

[10] Nie, Bruza and Lalmas

[11] The details in my book (see references)

[12] The well known probability ranking principle by S.E. Robertson.

The usual interpretation of the symbols is that H is a hypothesis (one of several) which is supported (or otherwise) by some evidence e . P(H|e) is interpreted as the probability that H is true given certain evidence e. Bayes' Theorem is a form of belief revision in that P(e|H), P(H), and P(e) are probabilities associated with propositions (or events) *before* e is actually observed and P(H|e) is the probability of H *after* e has been observed. The notation used to express this is especially opaque because the same P(.) is used for the prior and posterior probabilities. In fact P(H|e) would be better written as $P_e(H)$ indicating that we now have a *new* probability function P_e . A further difficulty with this approach is that the evidence e has to be certain at observation, that is, cannot be disputed once it is observed, which implies that $P_e(e) = 1$[13]. The fact that in general a conditional probability can only take into account an exact and certain proposition, or event, as its basis for revising the probability in the light of observation, is a source of some difficulty. In IR we use the Bayesian approach by calculating $P_q(R|d)$ through P(R|**x**) where **x** is some representation of the document assumed to be certain. This means that the description **x** is assumed to be true of the document, or *in* the document, depending on one's point of view. In many cases this is not fully appropriate, as the description, or representation **x** may be subject to uncertainty itself at the time of observation.

Let us examine the calculation of P(R|**x**) in a little more detail: let **x** be a set of independent variables x_1, ...x_n., then

[13] There is an ambiguity here but treat P_e as the revised probability.

$$P(A|G) = .4 \qquad P(A|B) = .4 \qquad P(A|V) = .8$$

Prior to inspection

$$\begin{aligned} P(A) &= P(A|G)\,P(G) + P(A|B)P(B) + P(A|V)P(V) \\ &= .4 \text{ x } .3 + .4 \text{ x } .3 + .8 \text{ x } .4 = .56 \end{aligned}$$

After inspection Jeffrey proposes:

$$\begin{aligned} P^*(A) &= P(A|G)P^*(G)+P(A|B)P^*(B)+P(A|V)P^*(V), \\ &= .4 \times .7 + .4 \times .25 + .8 \times .05 = .0485 \end{aligned}$$

known as *Jeffrey's rule of conditioning*

It is valid whenever $P^*(A|E_i) = P(A|E_i)$ where E_i is a partition of the sample space. This differs from Bayesian conditioning which would use $P^*(G) = 1$, or $P^*(B) = 1$, or $P^*(V) = 1$ and so revise $P(A)$ to $P^*(A) = P(A|X)$ when $X = G$, B, or V. Thus Bayesian conditioning can be seen as a special case of Jeffrey conditioning.

In this example the conditioning event is a simple property, namely colour, in what was proposed earlier, Δ—>q is a logical relationship. The machinery required to calculate the relevant probabilities is not so simple and I suggest you consult some of my papers. It is interesting that this complex expression defaults to some of the esarlier models under certain assumptions. Assume that Δ is empty then

$$P^*(Rel) = P(Rel \mid q)\, P^*(q) + P(Rel \mid \bar{q})\, P^*(\bar{q}).$$

In the classical case $P^*(q) = 1$ implies $P^*(Rel) = P(Rel \mid q)$ whereas $P^*(\bar{q}) = 1$ implies $P^*(Rel) = P(Rel \mid \bar{q})$. Boolean retrieval would judge that

$$\begin{aligned} P^*(q) = 1 &\Rightarrow P(Rel \mid q) = P^*(Rel) = 1 \\ P^*(q) = 0 &\Rightarrow P(Rel \mid \bar{q}) = P^*(Rel) = 0 \end{aligned}$$

References

P.D. Bruza, Stratified Information Disclosure: A synthesis between hypermedia and information retrieval, PhD thesis University of Nijmegen (1993).

D.J. Harper and A.D.M. Walker, ECLAIR: an extensible class library for information retrieval, *The Computer Journal*, **35**, 256-267 (1992).

R.C. Jeffrey, The Logic of Decision, 2nd Edition. Chicago: University of Chicago Press (1983)

M.Lalmas and C.J. van Rijsbergen, Alogical model of information retrieval based on situation theory, Proceedings 14th Information Retrieval Colloquium, Lancaster (1992)

M.E. Maron and J.L. Kuhns, On relevance, probabilistic indexing and retrieval, *Journal of the ACM*, **7**, 216-244 (1960).

J. Nie, Un modele de logique general pour les systemes de recherche d'informations. Application au prototype RIME. These University Joseph Fourier, Grenoble (1990)

S.E. Robertson, The probability ranking principle in IR. *Journal of Documentation*, **33**, 294-304 (1977).

C.J. van Rijsbergen, Information Retrieval, Second Edition. London:Butterworths (1979).

C.J. van Rijsbergen, A non-classical logic for Information Retrieval. *The Computer Journal*, **29**, 481-485 (1986).

C.J. van Rijsbergen. Towards an Information Logic. In N. Belkin and C.J. van Rijsbergen (eds.) *Proceedings of the Twelfth Annual ACM SIGIR Conference on Research and Development in Information Retrieval.* New York:ACM, 77-86 (1989).

C.J.van Rijsbergen, Probabilistic retrieval revisited, *The Computer Journal*, **35**, 291-298 (1992)

DISCUSSION

Rapporteur: Cecilia Calsavara

Lecture One

Dr Aalders asked Professor van Rijsbergen in the context of the problem of finding a needle in a hay stack, if the human approach of finding the needle would involve extra knowledge or extra information. Professor van Rijsbergen answered that it would involve extra knowledge.

Dr Larcombe enquired if instead of searching for a needle, one had been searching for a piece of hay, what was the importance of metrics in the process of searching. Professor van Rijsbergen answered that metrics were quite important depending on the space (context) considered. However, the metrics defined for one person may not reflect the needs of another person.

Professor Lincoln asked if the approach had been implemented in neural net. Professor van Rijsbergen said that it had been implemented and he had obtained very interesting results but he was very cautious because the results obtained for small systems may not be generalised for large applications.

Professor Wheeler asked if in the probabilistic model the attributes of the probabilistic function had to be independent. Professor van Rijsbergen answered that they did not need to be independent. Professor van Rijsbergen added that in one of the models that he had built the dependency between the attributes were second order. This led to a computational unsolvable solution and, in the end, the order was one and a half. He concluded saying that there are some interesting tradeoffs between the establishment of the dependency order of the attributes and the complexity of the system.