Using Conditional Logic for a More Complete Modelling of Information Retrieval

Jian-Yun Nie
Département d'informatique et de recherche opérationnelle
Université de Montréal
CP. 6128, succursale Centre-ville, Montreal, Quebec
H3C 3J7 Canada
nie@iro.umontreal.ca

Information retrieval (IR) aims to provide *relevant* information to a user's information need. Relevance, as shown by various cognitive and experimental studies, depends on a number of situational factors other than topic such as the user's prior knowledge, intentions, the state of available information, and so on. Formal IR models, on the other hand, only consider relevance from a topicality point of view. There is an urgent need of appropriate models which are able to consider relevance in a broader context. We suggest conditional logic as a potential tool for this purpose.

1. Introduction

Information retrieval (IR) aims at providing *relevant* information for an information need. However, at its current stage IR systems do not fit all our requirements. They often mistakenly retrieve irrelevant documents while missing relevant ones. This poor quality is strongly due to our current approach to information retrieval: an IR system is still widely viewed as a system that selects documents on the basis of keyword matching between document and query representations. This approach suffers from an important drawback: it is situation-independent, that is, the relevance estimation made by a system is independent of the user and of the context in which the retrieval is intended. As a consequence the current IR systems can only give relevance estimation based on topicality (subject) [5].

It has been known for a long time that topicality is not the only criterion of users' relevance judgments [14]. There are a number of other factors. For example, the user's prior knowledge on the subject, the expected use of information, the information already available, time constraint, and so on. We call the latter *situational* factors. Since late 1960s various studies have confirmed the diversity of situational factors from both cognitive and experimental points of view. Different terms have been used in literature to refer to these factors, such as *state of the user* [3], *situationality of the user* [4], *the problem state* [2, 15] and *information need situation* [1].

Although we now have many computational models for topical relevance estimation, there is no suitable model to deal with relevance in its full context. One of the reasons is the inability for the traditional tools to cope with various situational factors. In this article we will show that conditional logic provides an interesting alternative framework which may be suitably used for IR modeling in a broader context.

We will first point out the inability of some traditional modelling tools. Then a conditional logic approach is suggested.

2. Previous modeling

Existing computational IR models are restricted to topicality in their relevance estimation. This restriction is somewaht imposed by the modelling tools used. To see this, we will examine the two most used IR models, the Boolean model and the probabilistic model, with respect to some important situational factors.

The Boolean model

In this model, a document is represented by a set of terms. A term denotes the proposition "there is relevant information about a certain concept". A document representation then means that there is relevant information about a set of concepts. The set of terms representing a document is also commonly considered as a logical conjunction. A query is represented as a Boolean expression of terms, corresponding to the requirement that relevant documents about some concepts are needed.

Given such a document representation d and a query representation q, the estimation of the document's relevance is based on the following deduction:

$$\vdash d \supset q.$$

It is not possible to incorporate domain knowledge in this model. In order to build more "intelligent" IR systems, the simple model is extended such that domain knowledge can be modeled as a set S of logical sentences (implications). A document is relevant to a query with respect to S if we have the following deduction:

$$S \vdash d \supset q.$$

Although domain knowledge can be incorporated now, we still cannot cope with the changing situation-related factors such as the user's prior knowledge (which varies from user to user) and the available (or known) information (which varies from time to time). For example, if we want to consider the available information within S, the availability of some information may be expressed as a proposition in S. Suppose that in a given situation no information about "information retrieval" is already available to a user, we have: $\neg information\text{-}retrieval \in S$. Then given a document about "information retrieval" and any query representation q, the Boolean model will be able to deduce ($information\text{-}retrieval \supset q$) from S because the proposition *information-retrieval* is contradictory to S. This will lead to the wrong conclusion that the document is relevant to any query.

The same counterintuitive judgment occurs when we consider in S user-related factors such as the user's prior knowledge (which may be wrong). For example, a computer novice may believe that "World-Wide Wed" has nothing to do with "information retrieval" before retrieving any relevant document. Suppose now that a document describes exactly the relationship between the two concepts and the relationship is represented in the document representation d, then d is in contradiction with the user's prior knowledge (which is part of S). The Boolean model will also judge the document to be relevant to any query.

Although the above judgements are counterintuitive, they did not attract much attention in IR because IR modeling was concentrated on topicality. To model situational factors, the Boolean model becomes inappropriate.

Probabilistic IR models

In probabilistic models a document is represented by a set of (possibly weighted) terms. d represents the event that the corresponding document is retrieved. q represents the event that the corresponding query is satisfied by the relevant documents. Equivalently, q also represents the set of all the relevant documents to that query.

Given such d and q, most probabilistic IR models try to estimate relevance according to the conditional probability $P(q|d)$ (or a relevance *odd* based on it) [11]. This probability is not directly calculable. Bayes' rule is then used so that:

$$P(q|d) = \frac{P(d|q)\,P(q)}{P(d)}$$

in which $P(d|q)$ denotes the probability that the considered document is among the relevant documents for the query. The conditional probability $P(d|q)$ may be estimated using sample documents or previous relevance judgments of some documents [11]. To make this possible, however, the binary independence assumption is often made. This assumption has often been criticized. Indeed, under the assumption, human knowledge in the domain of application can hardly be incorporated, for example, knowledge such as "programming is part of computer science". To solve this problem, a more recent probabilistic model [17] makes use of a Bayesian inference network [10] in which

elements (e.g. terms) can be connected with each other by arcs denoting dependence relations. A Bayesian network in this approach is aimed to encode the domain knowledge as well as the composition of documents and queries. Relevance estimate is based on the calculation of $P(q|d)$ using probability spreading [10].

There have been a number of other probabilistic models proposed for IR. Most of them are extensions of the independent probabilistic model. All these models are based on the conditional probability $P(q|d)$.

In these models probability revision is crucial in order to adapt the system to a particular situation. One may argue that after revision, it is hopeful that the model is adapted to a situation. This is true, but the adaptation can be made only to a single situation. For different situations (e.g. different users or different states of available information), different models should be built. This indicates the difficulty for these models to consider various situational factors in a single model.

3. Outline of a more complete model of IR

IR as a communication process

The analogy between communication and IR has often been used in IR to consider relevance [14]. Let us look at a more recent definition of relevance in communication:

> An assumption is relevant in a context if and only if it has some contextual effect in that context. ([16], pp.122)

An "assumption" in IR is the retrieval of a document (some information becomes available or known), a contextual effect is that the query is better satisfied, or there is more information concerning the query. Similarly to this definition, relevance in IR may be understood as follows:

> A document is relevant to a query in a given situation if and only if it contributes to a better satisfaction of the query.

In order to measure the contribution of a query we need to estimate the hypothetical satisfiability of the query after the document content becomes known or the document content has been learnt. This approach implied the simulation of a learning process. To estimate the hypothetical probability, we have two main approaches based on the conditional probability $P(q|d)$ or the probability of conditional $p(d \Rightarrow q)$. It is generally agreed that the second probability corresponds better to our learning process. This, especially, when the item to be learnt comes into contradiction with the original ones in the situation. This latter case often occur in learning (belief revision [6, 7]). It also occurs in IR when situational factors are considered (for example, in the case of wrong user's prior knowledge). Thus we suggest to use conditional logic to consider relevance judgment in a broadened IR context ($\Rightarrow$ denotes the counterfactual conditional connective).

From the uncertainty principle to Ramsay test

In [12, 13] the following uncertainty principle has been proposed:

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

This principle has often been taken as corresponding to a general method for relevance estimation in IR. However, the nature of the connective $\rightarrow$ is not completely specified. It may be the counterfactual conditional connective, but also the material conditional connective $\supset$. Although this latter interpretation is not intended by [12, 13], it is

compatible with the above principle: given a data set (containing the available information and the user's knowledge), if a document content d is unknown, then for any query q there is no need to extend the data set because $d \supset q$ is true. However, we cannot consider the document as relevant in this case.

The correct way to estimate relevance that is intended in [12, 13] corresponds more closely to the Ramsay test:

> To evaluate a conditional, we first have to minimally extend our stock of beliefs so as to satisfy the antecedent; then the evaluation of the conditional is determined by that of the consequent relative to the extended stock of beliefs.

The nature of the conditional in Ramsay test is better determined: it is a counterfactual conditional.

Modeling

The first approach consists of the following direct modeling:

> Given a document and a query in a situation, represented as logical expressions d and q respectively; the document's relevance is approximated by $P(d \Rightarrow q)$.

With this modeling the connective '$\Rightarrow$' can capture many interesting properties of relevance [8] such as non-weakening-antecedent, non-transitivity, non-triviality. Let us give a simple counterexample for transitivity. Imagine a situation in which a user considers a document about "AI and logic" to be relevant to a query on "AI". In the same situation, the user may also consider a document on "AI" to be relevant to a query on "AI but not logic". However, a document on "AI and logic" cannot be relevant to a query on "AI but not logic". This counterexample denies transitivity for relevance.

However, some other properties of $\Rightarrow$ are undesirable for relevance. For example, the following property of '$\Rightarrow$' is not appealing:

$$\models d \wedge q \supset (d \Rightarrow q)$$

It means in IR: if the content of a document is already known and a query is already satisfied, then the document is relevant to the query. In other words, when a query is satisfied, all the available information becomes "relevant" to that query. No user will accept this judgment.

Another questionable property is the following:

$$\models d \Rightarrow \top$$

That is, any document is relevant to an *empty* information need (i.e. a requirement asking for nothing). In this case it is more reasonable to provide the user with no document rather than all the documents.

Yet another property is reflexivity:

$$\models d \Rightarrow d$$

This property may seem intuitive in IR. However, we cannot accept it as valid in the case where the user is only interested by new information. If the information contained in a document is already known to the user, this document is irrelevant for the user even if it concerns exactly the topic of the query. Thus this property is not always desirable for relevance.

All the above properties are in fact related to the state of available information. In order to capture the criterion concerning available information, we suggest to model relevance as follows:

Principle 1. Given d and q, d is relevant to q if $P(d \Rightarrow q) > P(q)$

It means: If in a hypothetical situation where a document is provided, the query is better satisfied than it is, then the document should be considered relevant.

Principle 2. Given d and q, d is relevant to q if $P(d \Rightarrow q) > P(\neg d \Rightarrow q)$

This alternative corresponds to situations where the user accepts known information as relevant. It means: a document is relevant if it is preferred to retrieve it than not to retrieve it.

The difference with the previous one lies in the meaning of "better" satisfaction of a query. In the first principle, it is compared with the current situation; in the second principle, it is compared with the situation in which the document content is completely unknown. We can say that this second estimate is based on the "absolute" information value of a document to the query; whereas the first estimate is based on the "new" information value of the document to the current situation. Both principles can find their supporting arguments in IR practice. They are adapted to different intentions of the user. If the user is only interested in new information, then the first principle is more suitable. If the user accepts redundant information (for example, if the user intends to find documents in order to confirm some known concepts), then the second principle may be applied.

The essence of two principles of relevance can be captured by the following expressions:

$$d \rightarrow_1 q =_{\text{def}} (d \Rightarrow q) \wedge \neg q$$

$$d \rightarrow_2 q =_{\text{def}} (d \Rightarrow q) \wedge \neg(\neg d \Rightarrow q)$$

Further study on these two connectives with respect to different systems of conditional logic can make clear the properties that they have. Their properties correspond exactly what we want to obtain in the two different situations: either consider the available information or not. Among the properties, we note that the following sentences are valid in both Stalnaker's and Lewis' systems [9] of conditional logic:

$\neg(A \rightarrow_1 \top)$, $\neg(A \rightarrow_1 \bot)$,

$(((A \rightarrow_1 B) \wedge (A \rightarrow_1 C)) \supset (A \rightarrow_1 (B \wedge C)))$

$((A \rightarrow_1 C) \wedge (B \rightarrow_1 C)) \supset ((A \vee B) \rightarrow_1 C)$,

$(((A \rightarrow_1 B) \wedge (A \rightarrow_1 C)) \supset (A \wedge B) \rightarrow_1 C))$

$\neg(A \rightarrow_2 \top)$, $\neg(A \rightarrow_2 \bot)$,

$A \rightarrow_2 A$,

$(((A \rightarrow_2 B) \wedge (A \rightarrow_2 C)) \supset (A \rightarrow_2 (B \wedge C)))$

But $\rightarrow_1$ and $\rightarrow_2$ do not have triviality, transitivity, weakening antecedent, contraposition, and $(A \wedge B) \supset A \rightarrow B$.

All these properties have adequate interpretation in IR. It is thus suggested that conditional logic is a promising modelling tool for IR in a broadened context.

References

1. C. L. Barry, User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science*, vol. 45, pp. 149-159 (1994).
2. N. J. Belkin and A. Vickery, *Interaction in information systems*. West Yorkshire, UK: The British Library Board (1985).
3. C. A. Cuadra and R. V. Katter, Opening the black box of "relevance". *Journal of Documentation*, vol. 23, pp. 291-303 (1967).
4. B. Dervin and M. Nilan, Information needs and uses. *Annual Review of Science and Technology*, vol. 21, pp. 3-27 (1986).
5. T. J. Froehlich, Relevance reconsidered - Towards an agenda for the 21st century: Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science*, vol. 45, pp. 124-134 (1994).

6. P. Gärdenfors, Imaging and conditionalization. *The Journal of Philosophy*, vol. 79, pp. 747-760 (1982).
7. P. Gärdenfors, *Knowledge in Flux: Modeling in Dynamics of Epistemic States*. Cambridge: MIT Press (1988).
8. J.-Y. Nie, F. Lepage, and M. Brisebois, Information retrieval as counterfactual. *The Computer Journal* (to appear).
9. D. Nute, *Topics in Conditional Logic*, vol. 20. Dordrecht, Holland: Reidel (1980).
10. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo CA: Morgan Kaufmann (1988).
11. C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworths (1979).
12. C. J. van Rijsbergen, A non-classical logic for information retrieval. *The Computer Journal*, vol. 29, pp. 481-485 (1986).
13. C. J. v. Rijsbergen, Towards an information logic. *Research and Development on Information Retrieval - ACM-SIGIR*, 77-86 (1989).
14. T. Saracevic, The concept of "relevance" in information science: a historical review. in *Introduction to Information Science*, T. Saracevic, Ed. New York: R. R. Bowker Company, pp. 111-151 (1970).
15. T. Saracevic *et al*., A Study of information seeking and retrieving (I): background and methodology. *Journal of the American Society for Information Science*, vol. 39, pp. 161-176 (1988).
16. D. Sperber and D. Wilson, *Relevance: Communication and cognition*. Oxford: Basil Blackwell (1986).
17. H. Turtle and W. B. Croft, Inference network for document retrieval. *Research and Development on Information Retrieval - ACM-SIGIR*, Brussels(1990).