Towards a Logical Reconstruction of Information Retrieval Theory

Fabrizio Sebastiani
Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
Via S. Maria, 46 – Pisa, Italy
E-mail: fabrizio@iei.pi.cnr.it

Abstract. We here expand on a previous paper concerning the role of logic in information retrieval (IR) modelling. In that paper, among other things, we had pointed out how different ways of understanding the contribution of logic to IR have sprung from the (always unstated) adherence to either the *total* or the *partial knowledge assumption*. Here we make our analysis more precise by relating this dichotomy to the notion of *vividness*, as used in knowledge representation, and to another dichotomy which has had a profound influence in database theory, namely the distinction between the *proof-theoretic* and the *model-theoretic* views of a database, spelled out by Reiter in his "logical reconstruction of database theory". We show that precisely the same distinction can be applied to logical models of IR developed so far. The strengths and weaknesses of the adoption of either approach in logical models of IR are discussed.

1 Introduction

Logical models of information retrieval have been actively investigated in the last ten years. The reason behind this interest in logic on the part of IR theorists springs from a substantial dissatisfaction with the insights into the very nature of information, information content, and relevance, that mainstream IR research gives. A few years ago William Cooper, one of the major researchers in the history of information retrieval, sharply summarised the status of IR theory by saying that "(...) deep down inside it's shallow" [2, page 201]. And the fact that things have not changed since then is witnessed by Alan Smeaton's recent comment on logical modelling of IR: "(...) it is without doubt that if there is ever to be a really significant breakthrough in information retrieval, it will come from this kind of fundamental and basic work" [35, page 13].

We indeed think that there are two main IR-related issues to which logic might provide better answers than current approaches.

The first issue has to do with the *quantitative* view of information content that the received wisdom of IR embodies. According to this view the degree of similarity (or the probability of relevance, depending on the adopted model) of a document to a given request may be estimated, by and large, by computing the occurrence frequency of words in the request, in the document candidate for retrieval, and in the collection of documents being searched[1]. If on one hand these quantitative methods are still unsurpassed (see e.g. [11]) in terms of effectiveness (i.e. in terms of their ability to weed the irrelevant documents from the relevant ones), on the other hand they do not constitute, in all evidence, a satisfactory explanation of the fundamental notions of IR. In other words, it is implausible that the very notion of information content of a document may ultimately come down to word counts, irrespective of the syntactic, semantic and pragmatic role that each individual word occurrence plays; the dominant view in linguistic semantics is that information content *must* be more than that, even if we are not yet able to put our fingers on it. Quantitative

[1] For instance, the fact that a term occurs frequently in a document adds weight to the hypothesis that this term is "important" to the document meaning; conversely, the fact that a term appears frequently in the document collection subtracts weight to the hypothesis that this term is important to the meaning of individual documents in which it occurs.

models thus provide a *phenomenology*, rather than a theory, of information content and relevance[2]. It is exactly the search for a *theory* that has driven many IR "theorists" to the investigation of logical models of IR. Far from believing that the ultimate IR system will be a theorem prover, many of these investigators are convinced that logic, by its strong reliance on the semantic aspect of the formulae it deals with, may foster our understanding of the fundamental (and inherently semantic) notions of information, information content and relevance.

The second issue has to do with the separation of concerns that current practice in IR has *de facto* established between the issues of 1) *representing* the content of documents and requests (*indexing*), and 2) *reasoning* with such representations in order to establish the relevance of the former to the latter (*matching*). In fact, in present-day IR, indexing techniques are only loosely bound to the matching techniques that use the representations built by them; for instance, the same method for computing the representations of documents/requests (e.g. $tf * idf$ weighting) is being used in conjunction with widely different matching techniques, and the same matching technique (e.g. the cosine measure) is being used in conjunction with representations of documents/requests obtained by widely different methods (see [29] for an example of this "combinatoric" coupling of indexing and matching). In logic, IR theorists find instead a framework in which representation and reasoning are not independently motivated, but are, in some sense, one and the same thing. Logic prompts IR theorists not only to clearly specify the semantics of the representation language for documents/requests and the semantics of relevance, but also to ensure that the way actual representations are arrived at is consistent with this semantic specification!

In a previous paper [34] we analysed the literature on logical models of IR from the point of view of their compliance with the well-formedness criteria that are standard in applied logic. In [34, Section 6.1] we argued that, from this literature, two different ways of understanding the contribution of logic to IR modelling emerge, and that each of them is based on the (unstated) adoption of either the *total knowledge* or the *partial knowledge* assumption.

At a first approximation, the total knowledge assumption means that *everything about the problem domain is assumed to be known*. Although this characterisation may look a bit strong at first sight, it is not once one interprets "everything" as "everything that can be stated in the logical language used for the representation of the problem domain". For instance, the traditional Boolean model of IR (in which the logical language for the representation of documents is that of Boolean conjunctions of propositional letters) is a model in which total knowledge is implicitly assumed. To see this, assume that the set of propositional letters (i.e. the controlled indexing language) is $\mathcal{L} = \{t_1, \ldots, t_n\}$. The key observation is that, if a document d_i is represented e.g. by the conjunction $t_2 \wedge t_5 \wedge t_7$, this is assumed to mean not only that d_i is about t_2, t_5 and t_7, but also that d_i is *not* about t_i for $i \neq 2, 5, 7$. In other words, the truth value of *everything that can be specified in the language about* d_i (i.e. whether, for a given t_j, d_i is or is not about t_j) *is assumed known*.

The total knowledge assumption is present, although better hidden, also in some of the models that make use of formal tools traditionally viewed as means of representing uncertainty (even though uncertainty is closely associated to the notion of the partiality of knowledge!). For instance, in the *extended Boolean model* of [30] or in the *probabilistic model* of [26], documents are represented as conjunctions of *weighted* terms, where the weight $w_{ij} \in [0, 1]$ of term i in document j is taken to represent the "discrimination power" of the term [31] or its probability of occurrence in a generic relevant document [26]. Here, the key observation is that the weight of any term, whatever its interpretation, *is always assumed known*. In other words, the presence of uncertainty in these models is, in the precise sense exposed above, only apparent: every sentence that can be expressed in the representation language is either known to be true or known to be false.

The partial knowledge assumption, instead, makes explicit the fact that not all that is representable in the chosen logical language is assumed to be known. For instance, one may conceive a variant of the Boolean model of IR in which, given the usual set of propositional letters

[2] Word frequency distributions constitute a rich source of information for text analysis, and are exploited not only in information retrieval but also in quantitative stylistics, lexicography, and linguistics; see e.g. [1] for an example of such applications.

$\mathcal{L} = \{t_1, \ldots, t_n\}$, a document representation $t_2 \wedge t_5 \wedge t_7$ for document d_i is taken to mean, among other things, that *it is not known* whether d_i is or is not about t_j for $j \neq 2, 5, 7$. In other words, models relying on the partial knowledge assumption do not take commitments concerning what is not explicitly represented. Adherence to the partial knowledge assumption entails reasoning in the presence of incomplete information, which is the standard way of performing inference in logic.

For what we have said up to now (and, for that matter, for what we had said in [34, Section 6.1]), the distinction between total-knowledge and partial-knowledge models of IR might as well come down to the better known distinction between adopting either a *closed world assumption* (CWA) or an *open world assumption*, respectively (see e.g. [17, Chapter 7]). In this paper we argue that our distinction amounts to more than that, in that total-knowledge models of IR assume not only that everything about the problem domain is known, but also that it is represented in *vivid* form [13, 14]. The consequence is that adopting either the total-knowledge or the partial-knowledge assumption means taking an implicit stand as to what, logically speaking, a (representation of a) document collection is: more precisely, the two different positions relate to the dichotomy between the *model-theoretic* and the *proof-theoretic* models of databases, exposed in a paper by Reiter [24]. The aim of this paper is to show how this latter dichotomy may usefully be applied to the case of logical models of IR, and how advantages and disadvantages of either approach that have already been discussed in the DB literature apply, and to what extent, to the IR case.

The paper is structured as follows. In Section 2 we briefly discuss the notion of "vividness", and how it relates to the model-theoretic and the proof-theoretic views of databases. In Section 3 we discuss how this dichotomy applies to IR too, show how proposed logical models of IR have *de facto* adhered to either camp, and discuss the advantages and disadvantages that these models incur into by way of this adherence. Section 4 discusses the issue of how "model-theoretic" models may be recasted in proof-theoretic form. Section 5 concludes.

2 Vividness, the model-theoretic, and the proof-theoretic models of databases

2.1 Vivid knowledge bases and total-knowledge models of IR

Implicit in the discussion of the previous section is the fact that partial-knowledge models implicitly assume that the problem knowledge *cannot* be encoded by means of a *complete theory* of the chosen logic[3], and therefore tend to rely on reasoning methodologies involving deduction (possibly of a probabilistic kind) that make use of a knowledge base representing the incomplete theory (see e.g. [18, 32]). Such a theory has more than one model, and deduction may as usual be seen as a compact way of handling them all.

Total-knowledge models (see e.g. [4, 5, 6, 21, 22]) are instead built along the assumption that the problem knowledge can be encoded by means of a *complete theory* of the chosen logic. However, a key point that we missed to observe in [34, Section 6.1] is that these models assume that, of this complete theory, *the simplest representation possible is always available*, where this representation is what Levesque [13, 14] calls a *vivid* knowledge base, i.e. a set of (possibly negated) *ground*, *atomic* statements. An example of a vivid knowledge base for the language of propositional logic built upon the alphabet $\mathcal{T} = \{t_1, t_2\}$ is the set $KB = \{t_1, \neg t_2\}$: it is a complete theory (its only satisfying interpretation is the truth value assignment that assigns T to t_1 and F to t_2), and it also vivid, unlike e.g. its equivalent knowledge base $KB' = \{\neg(t_1 \supset t_2)\}$. Levesque observes that, when a knowledge base is in vivid form, it basically consists in an "analogue" of its unique satisfying interpretation, and therefore may be reasoned upon by methods quicker that theorem proving, much in the same way in which a photograph of a tree in front of a house immediately allows us

[3] In logic, a *complete theory* is a deductively closed set of formulae Γ such that, for any formula α in the language, either $\alpha \in \Gamma$ or $\neg\alpha \in \Gamma$. An *incomplete theory* is a deductively closed set of formulae for which this property does not hold. Therefore, a (consistent) theory has a unique satisfying interpretation iff it is complete, while it has more than one satisfying interpretation iff it is incomplete.

to reach the conclusion, with no complex chains of either disjunctive or implicational reasoning, that there is a tree in front of a house[4].

It is exactly the vividness of the representations upon which total-knowledge models are built that allows them to disregard proof theory and theorem proving, favouring instead approaches to document relevance estimation based on the explicit manipulation of the vivid data structure that represents the complete theory. For instance, in the "imaging" models of [4, 5, 6] documents are represented as in the Boolean model, but the problem domain is additionally represented by a probability density function $\mu(t)$ on the set of n terms occurring in the document collection (again, taken to represent the "importance" of the term relative to other terms in the collection), and by a real-valued function $\sigma(t_1, t_2)$ on the set of pairs of terms (taken to represent the "semantic relatedness" between the two terms). Here, not only the "importance" of all terms is always assumed known, but it is represented *explicitly* as a vector of weights of length n; not only the "semantic relatedness" between any two terms is always assumed known, but it is represented *explicitly* as an $n \times n$ bidimensional matrix of weights. Should any of these items of knowledge be not explicitly available, and therefore be inferred on demand from the application of inference rules to other items of knowledge, the vividness property would be lost, and the methods for reasoning on vivid representations would no more be applicable[5].

2.2 Vivid knowledge bases and data bases

Levesque [13] observed that a vivid knowledge base, being free from disjunctive, implicational or quantified knowledge, is akin to a relational database, where reasoning is basically achieved simply by the lookup of the required information in a table where all available information is stored in ready-to-use form. This suggests the existence of a connection between total-knowledge logical models of IR and databases. This connection may be better appreciate in the context of the distinction between the *proof-theoretic* and the *model-theoretic* view of databases exposed by Reiter in his seminal paper [24]. According to

1. the model-theoretic view, a database is an interpretation I of a first-order logical language L, a query is a formula α of L, and query evaluation may be seen in terms of checking the truth of α in I;
2. the proof-theoretic view, a database is a set Γ of formulae of a first-order logical language L, a query is a formula α of L, and query evaluation may be seen in terms of proving that α belongs to the deductive closure of Γ.

Reiter argues that the proof-theoretic view of databases is more fruitful than the model-theoretic one. The latter, in fact, shows its limits in the impossibility of dealing with incomplete knowledge (since first-order interpretations are complete specifications of a state of affairs), null values (since no "undefined" truth value is catered for by first order semantics), and, above all, domain knowledge. In the next section we will discuss how these issues impact on logical models of information retrieval.

[4] To take propositional logic as an example, one may observe that deciding whether α logically follows from a knowledge base Γ in vivid form may be achieved by checking whether Γ is a satisfying truth assignment for α, a substantially easier task that doing unrestricted theorem proving. The careful reader will remember this observation when reading Footnote 9.

[5] Interestingly enough, an analysis of the major traditional (non-logical) models of IR (e.g. [27, 30, 31]) reveals that the total knowledge assumption (or its non-logical equivalent) seems to be "wired" into information retrieval since its very inception, no doubt because of its greater computational tractability. This does not mean that the designers of IR models or systems are unaware of the fact that the basic quantities of IR, such as the "importance" of a term, cannot be determined with certainty; it means that the systems (viewed as cognitive agents) are!

3 Model-theoretic and proof-theoretic models of information retrieval

The very idea of a logical model of IR, put forth by van Rijsbergen in [39], relies on the estimation of the formula $P(d \rightarrow r)$, where d and r are logical formulae representing the document under consideration and the request, respectively, $P(\alpha)$ stands for "the probability of α", and $\rightarrow$ is the conditional connective of the logic in question. As discussed in [34], this proposal has been interpreted by researchers to mean widely different things, and has originated two broad classes of models, total-knowledge models and partial-knowledge models.

It is the contention of this paper that the difference between total-knowledge and partial-knowledge models may exactly be seen in terms of Reiter's distinction between viewing databases model-theoretically or proof-theoretically.

Let us discuss total-knowledge models first. It suffices to analyse any model in this category (we will use as examples the "imaging" model developed in [5]) to recognise the basic traits of Reiter's model-theoretic view:

- the reliance on a logic $\mathcal{L}$ endowed with a model-theoretic semantics for its language L. In [5], $\mathcal{L}$ is the **C2** conditional logic, L is the language of propositional letters[6], and the semantics is a model-theoretic semantics based on possible worlds and the "imaging" principle (see e.g. [15]);
- the "representation" of the data encoding the problem domain (i.e. documents, requests, terms, ...) not by the exclusive means of formulae of L, but also by means of a data structure representing a semantic interpretation I of L, similarly to Point (1) in Section 2.2. In [5] documents (and requests) are represented by propositional letters, terms are represented by possible worlds (on which a probability density function $\mu(t)$ is defined, representing the importance of the term relative to other terms in the collection) and the semantic relatedness between terms is represented by a real valued function $\sigma(t_1, t_2)$;
- a reasoning method not aimed at determining validity in $\mathcal{L}$, but aimed instead at determining truth in the unique satisfying interpretation I, usually by the explicit manipulation of I itself[7]. In [5], $P(d \rightarrow r)$ is computed by revising $\mu(t)$ in a d- and σ-dependent way to yield $\mu'(t)$, and to subsequently compute $P(r)$ on $\mu'(t)$; no use of the proof theory of **C2** is made.

Partial-knowledge models follow instead not only the fundamental traits of Reiter's proof-theoretic view of databases, but also the standard guidelines of applied AI-style knowledge representation. We will take as example the model presented in [32] to illustrate the following basic features:

- the reliance on a logic $\mathcal{L}$ endowed with a model-theoretic semantics for its language L. In [32], $\mathcal{L}$ is the $\mathcal{P}$-MIRTL probabilistic description logic, L is its language of "concepts" and "roles", and the semantics is, again, a model-theoretic semantics based on possible worlds;
- the representation of the data encoding the problem domain by the exclusive means of formulae of L, similarly to Point (2) in Section 2.2. In [32] documents, requests and terms are represented by concepts, and their relative "importance" is represented by qualifying concepts probabilistically;
- a reasoning method aimed at determining validity in $\mathcal{L}$, i.e. truth in the many interpretations satisfying the representation of the problem domain. In [32], $P(d \rightarrow r)$ is computed by finding the real number $v \in [0, 1]$ for which $P(d \rightarrow r) = v$ is valid in the theory representing the problem domain[8].

[6] The **C2** logic was originally defined on a full propositional language [37].

[7] The total knowledge assumption is so widespread in information retrieval (see Footnote 5) that these two notions are often collapsed in IR models: Wong and Yao [41], for instance, state that "The notion of relevance in the Boolean model is interpreted as a strict logical implication: a document is retrieved only if it logically satisfies a request." See [34] for a thorough discussion of this point.

[8] The Datalog-inspired approach of Fuhr [7] is a particular case of the proof-theoretic approach, because its semantics, being informed by the closed world assumption, is such that the theory that represents the problem domain is complete, as in the model-theoretic approach. In Reiter's scheme, the approach of [7] would thus be classified as proof-theoretic with *absence* of incomplete information, while the above-discussed approach of [32] would be labelled proof-theoretic with *presence* of incomplete information. A position similar to the one of [7] is adopted in [19].

At this point, a discussion of the relative advantages and disadvantages of the two opposing views is in order.

The advantage that accrues from the adoption of the model-theoretic perspective is of a "berry-picking" nature: rather than exploiting *logic* (and the tools provided by meta-logic such as the notions of logical consequence, validity, and the like), one picks *one particular intuition* embodied by *one particular logic* and applies it in what is essentially an extra-logical context. This is the case of [5], that, rather than exploiting the inferential power of the **C2** conditional logic in performing inference, borrows the particular graph-theoretic topology of **C2**'s semantic structures and applies it to the revision of a probability density function for establishing relevance. This is also the case of [21], that equates the distance that separates a document from "perfect" relevance to a request, to the distance that separates the nodes representing the document and the request, respectively, in a graph that resembles the "Kripke structures" used for giving semantics to modal logic. In not making use of proof-theory (and, hence, of inference) these approaches exploit the underlying total knowledge assumption, thus avoiding the added computational burden that the existence of multiple interpretations, which would accrue from the partiality of knowledge, brings about.

Another advantage that should be mentioned is the fact that insights from more traditional (non-logical) models of IR may be incorporated in a total-knowledge model without effort, as these other models are also based on the total knowledge assumption. For instance, the *idf* measure of the discrimination power of terms may be incorporated in any total-knowledge model that requires relative term importance to be measured, as both the former and latter models are based on the common premise that the "weight" of a given term, whatever its interpretation, is always known.

The advantages deriving from the adoption of a proof-theoretic perspective, instead, are due to the fact that this perspective opens the way to the *exploitation of domain knowledge in establishing relevance*. As explained by Reiter [24, page 193], the very possibility of a proof-theoretic view of DBs "by itself (...) would not be a very exciting result. (...) The idea bears fruit only in its capacity for generalization". And the usefulness of opening up the retrieval process to the incorporation of additional sources of information is amply recognized also from the IR community: Wong and Yao [41, page 41], for instance, champion the adoption of the "subjective" view of probability also on the grounds that "it provides an effective means to incorporate semantic information into the retrieval process".

We think that the importance of incorporating domain knowledge becomes especially evident once one considers that the knowledge that should be brought to bear in the retrieval process may either be

- *endogenous*, i.e. with an internal origin. This is the case of all types of knowledge that can be *estimated* (rather than computed deterministically) through a process of automatic information extraction from the document or from the document collection. Examples of this are the discriminating power of a term (in textual retrieval), the shapes of objects portrayed in photographs (in image retrieval), or the individual words pronounced by speakers (in speech retrieval);
- *exogenous*, i.e. with an external origin. This is the case of all types of knowledge that, either inherently or due to the limitations of current technology, cannot be extracted automatically, but have to be provided "manually", i.e. from an external source; examples of this are the author of a photograph (in image retrieval) or the nationality of a non-native speaker (in speech retrieval).

Traditional information retrieval research, from Luhn [16] onwards, has assumed that retrieval should be based on endogenous knowledge only. Today, this assumption is increasingly challenged by the emergence of novel applications such digital libraries and multimedia search engines, and by the increasing convergence of research fields that had traditionally led a separate existence, such as IR, DBs, and on-line library catalogues. In these newer contexts, the integration of different sources of knowledge is essential. In order to achieve the level of effectiveness that nowadays users demand, *resource discovery*, the reincarnation of IR in the open-ended context of digital libraries, and *multimedia document retrieval*, cannot rely exclusively on endogenous knowledge, but need

to be supported by additional, exogenous information, supplied either by the authors themselves (e.g. under HTML "META" tags), or by third-party cataloguers. To address requests such as

black and white photographs of successful actors of silent movies

an IR system should rely both on endogenous knowledge (knowing whether the image is a black and white one; (maybe) knowing whether a person is portrayed) and exogenous knowledge (knowing whether the portrayed person was an actor; whether he or she was successful; whether he or she has actually played in silent movies).

To allow the integration of exogenous and endogenous knowledge, a proof-theoretic approach is essential, as its fundamental assumption that knowledge is incomplete allows different sources of knowledge to be smoothly integrated, by simply adding together the corresponding sets of formulae in an incremental fashion.

4 From model- to proof-theory

Is it possible to recast in proof-theoretic terms a model of IR originally designed along model-theoretic guidelines, and viceversa? Is it worthwhile? An answer to the latter question is implicit in our discussion of the "berry-picking" advantages of the model-theoretic approach and of the "exogenous knowledge" advantages of the proof-theoretic one: one might want to embody in one's model the intuitions coming from the model-theoretic semantics of a given logic (e.g. the "imaging" principle exploited in [5]) and, at the same time, want to incorporate exogenous knowledge in the model. The former question then becomes crucial. It should be clear that the first option (i.e. model $\rightarrow$ proof) is possible, while the latter (i.e. proof $\rightarrow$ model) is not, the reason being that total knowledge is just a particular case of partial knowledge, but not vice-versa; a situation in which the knowledge of the domain is only partial is far more complex from the reasoning viewpoint, and this is well reflected in the smaller computational complexity of model checking with respect to theorem proving[9].

In the case of DBs, the feasibility of the "model $\rightarrow$ proof" option is well shown by Reiter [24], who describes a mapping from an interpretation I of a first order language to a first order theory Γ such that I and Γ provide model-theoretic and proof-theoretic characterisations, respectively, of the same database. Reiter's move is well-known in logic (although the relationship was apparently not noticed by Reiter himself), as it is an instance of what is called a *standard translation*. In general, a standard translation may be seen as the representation of the model theory of a logic $\mathcal{L}$ in the language (hence, in the proof theory) of a logic $\mathcal{L}' \neq \mathcal{L}$[10]. In the IR literature, a mapping conceptually similar to Reiter's (again, the relationship with standard translation, and with Reiter's work, was not noticed by the authors) is present in [3, 28] and [33], each dealing with recasting the **C2**-based "imaging" models of [4, 5, 6] in terms of a probabilistic logic (Fuhr's Probabilistic Datalog [7] in the case of [3, 28], Halpern's $\mathcal{L}_3$ logic [9] in the case of [33]).

As shown in [24], in order for these mappings to be faithful, it is generally necessary to introduce various axioms whose aim is to restrict the number of satisfying interpretations of the resulting theory to just one, i.e. the interpretation I from which the whole process began[11]. Similarly to the Reiter case, [33] introduces the following (1) domain closure axioms (saying that the only existing individuals are those referred to from within the database), (2-3) unique names axioms (saying that different individual constants refer to different individuals), and (4) completion axioms (saying

[9] This is basically the same difference between (a) a problem in P and (b) one in NP, as these may be characterized as (a) one in which a solution may be found polynomially, and (b) one in which a candidate solution may be checked polynomially [8]. Check also [10].

[10] The first and most famous example is the standard translation of modal logic into first order logic, proposed by van Benthem [38]; a general framework for doing standard translations is presented in [23].

[11] Some of these axioms are unneeded, and hence not introduced, in [3, 28], as they are already "wired" in the logical language (Probabilistic Datalog) onto which the mapping is performed.

that the only individuals that enjoy a given property are those for which this property is explicitly predicated).

$$
\begin{aligned}
& Term(t_1) \ \wedge \ldots \wedge \ Term(t_n) \\
& Doc(d_1) \ \wedge \ldots \wedge \ Doc(d_m) \\
& \forall x.[x = t_1 \vee \ldots \vee x = t_n \vee x = d_1 \vee \ldots \vee x = d_m] && (1) \\
& t_1 \neq t_2 \wedge t_1 \neq t_3 \wedge \ldots \wedge t_{n-1} \neq t_n && (2) \\
& d_1 \neq d_2 \wedge d_1 \neq d_3 \wedge \ldots \wedge d_{m-1} \neq d_m && (3) \\
& \forall x.\neg(Document(x) \wedge Term(x)) && (4)
\end{aligned}
$$

Interesting to our purposes is to note that by removing (or "typing") one or more of these classes of axioms from the theory, one may let exogenous knowledge in. For instance, typing our domain closure axioms means substituting (1) with (5) and/or (6):

$$
\begin{aligned}
& \forall x.Doc(x) \supset (x = d_1 \vee \ldots \vee x = d_m) && (5) \\
& \forall x.Term(x) \supset (x = t_1 \vee \ldots \vee x = t_n) && (6)
\end{aligned}
$$

which would allow other non-term and non-document entities (i.e. authors) to be talked about, thus enabling exogenous knowledge to be plugged in.

5 Concluding remarks

In this paper we have elaborated on one of the findings of [34], arguing that the distinction between total-knowledge and partial-knowledge models of IR may more fruitfully be interpreted in terms of Levesque's notion of vividness and Reiter's distinction between the model-theoretic and the proof-theoretic models of databases. This finding has several implications, especially for the possibility of incorporating knowledge originating from different sources into information retrieval systems, a necessity rather than a possibility in advanced information seeking environments such as multimedia document retrieval systems.

Quite independently of the practical impact on these advanced applications, we think that the present findings contribute in shedding light on some of the current theorizing on IR. To quote Robertson, we need to spell out the assumptions that underlie systems and models

> "not because mathematics *per se* is necessarily a Good Thing, but because the setting up of a mathematical model generally presupposes a careful formal analysis of the problem and specification of the assumptions, and explicit formulation of the way in which the model depends on the assumptions. (...) It is only the formalization of the assumptions and their consequences that will enable us to develop better theories." [25, page 128]

Acknowledgements

Past discussions with Gianni Amati, Iain Campbell, Fabio Crestani, Carlo Meghini, Jian-Yun Nie, Ulrich Thiel and Keith van Rijsbergen on the subjects of this paper have influenced the positions exposed here, even though often dissenting from them; I am grateful to all of them for the feedback they have given me. I am also grateful to Roberto Sebastiani for providing valuable pointers to the theorem proving literature.

References

1. R. Harald Baayen and Rochelle Lieber. Word frequency distributions and lexical semantics. *Computers and the Humanities*, 30:281–291, 1997.

2. William S. Cooper. Gedanken experimentation: an alternative to traditional system testing? In Karen Sparck Jones, editor, *Information retrieval experiment*, pages 199–209. Butterworths, London, UK, 1981.
3. Fabio Crestani and Thomas Rölleke. Issues in the implementation of general imaging on top of Probabilistic Datalog. In Mounia Lalmas, editor, *Proceedings of the 1st International Workshop on Logic and Uncertainty in Information Retrieval*, Glasgow, UK, 1995.
4. Fabio Crestani, Fabrizio Sebastiani, and Cornelis J. van Rijsbergen. Imaging and information retrieval: Variations on a theme. In Fabio Crestani and Mounia Lalmas, editors, *Proceedings of the 2nd International Workshop on Logic and Uncertainty in Information Retrieval*, pages 48–49, Glasgow, UK, 1996.
5. Fabio Crestani and Cornelis J. van Rijsbergen. Information retrieval by logical imaging. *Journal of Documentation*, 51:3–17, 1995.
6. Fabio Crestani and Cornelis J. van Rijsbergen. Probability kinematics in information retrieval. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, pages 291–299, Seattle, US, 1995.
7. Norbert Fuhr. Probabilistic Datalog: a logic for powerful retrieval methods. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, pages 282–290, Seattle, US, 1995.
8. Michael R. Garey and David S. Johnson. *Computers and intractability. A guide to the theory of NP-completeness*. Freeman, New York, US, 1979.
9. Joseph Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.
10. Joseph Y. Halpern and Moshe Y. Vardi. Model checking vs. theorem proving: a manifesto. In *Proceedings of KR-91, 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 325–334, Cambridge, US, 1991.
11. Donna Harman. Overview of the second Text REtrieval Conference (TREC-2). *Information Processing and Management*, 31:271–289, 1995.
12. William L. Harper, Robert Stalnaker, and Glenn Pearce, editors. *Ifs. Conditionals, belief, decision, chance and time*. Reidel, Dordrecht, NL, 1981.
13. Hector J. Levesque. Making believers out of computers. *Artificial Intelligence*, 30:81–108, 1986. Also reprinted in [20], pp. 69–82.
14. Hector J. Levesque. Logic and the complexity of reasoning. *Journal of Philosophical Logic*, 17:355–389, 1988.
15. David K. Lewis. Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, 85:297–315, 1976. Also reprinted in [12], pp. 129–147.
16. Hans P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1:309–317, 1957.
17. Witold Łukaszewicz. *Nonmonotonic reasoning: formalization of commonsense reasoning*. Ellis Horwood, Chichester, UK, 1990.
18. Carlo Meghini, Fabrizio Sebastiani, Umberto Straccia, and Costantino Thanos. A model of information retrieval based on a terminological logic. In Robert Korfhage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 298–307, Pittsburgh, US, 1993. Published by ACM Press, Baltimore, US.
19. Carlo Meghini and Umberto Straccia. A relevance description logic for information retrieval. In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pages 197–205, Zürich, CH, 1996.
20. John Mylopoulos and Michael L. Brodie, editors. *Readings in artificial intelligence and databases*. Morgan Kaufmann, San Mateo, US, 1989.
21. Jian-Yun Nie. An information retrieval model based on modal logic. *Information Processing and Management*, 25:477–491, 1989.
22. Jian-Yun Nie. Towards a probabilistic modal logic for semantic-based information retrieval. In *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval*, pages 140–151, Kobenhavn, DK, 1992.
23. Hans J. Ohlbach. Semantics-based translation methods for modal logics. *Journal of Logic and Computation*, 1(5):691–746, 1991.
24. Raymond Reiter. Towards a logical reconstruction of relational database theory. In Michael L. Brodie, John Mylopoulos, and Joachim W. Schmidt, editors, *On conceptual modelling*, pages 191–233.

Springer, Heidelberg, DE, 1984. Also reprinted in [20], pp. 301–326.
25. Stephen E. Robertson. Theories and models in information retrieval. *Journal of Documentation*, 33:126–148, 1977.
26. Stephen E. Robertson, M.E. Maron, and William S. Cooper. Probability of relevance: a unification of two competing models for document retrieval. *Information technology: research and development*, 1:1–21, 1982.
27. Stephen E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976. Also reprinted in [40], pp. 143–160.
28. Thomas Rölleke. Does Probabilistic Datalog meet the requirements of imaging? In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, page 374, Seattle, US, 1995.
29. Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing and management*, 24:513–523, 1988. Also reprinted in [36], pp. 323–328.
30. Gerard Salton, Edward A. Fox, and Harry Wu. Extended Boolean information retrieval. *Communications of the ACM*, 26(12):1022–1036, 1983.
31. Gerard Salton, Anita Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975. Also reprinted in [36], pp. 273–280.
32. Fabrizio Sebastiani. A probabilistic terminological logic for modelling information retrieval. In W. Bruce Croft and Cornelis J. van Rijsbergen, editors, *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 122–130, Dublin, IE, 1994. Published by Springer Verlag, Heidelberg, DE.
33. Fabrizio Sebastiani. Information retrieval, imaging and probabilistic logic. *Computers and Artificial Intelligence*, 17(1):35–50, 1998.
34. Fabrizio Sebastiani. On the role of logic in information retrieval. *Information Processing and Management*, 34(1):1–18, 1998.
35. Alan F. Smeaton. An overview of information retrieval. In Maristella Agosti and Alan F. Smeaton, editors, *Information Retrieval and Hypertext*, pages 3–25. Kluwer Academic Publishers, Dordrecht, NL, 1997.
36. Karen Sparck Jones and Peter Willett, editors. *Readings in information retrieval*. Morgan Kaufmann, San Mateo, US, 1997.
37. Robert C. Stalnaker and Richmond H. Thomason. A semantical analysis of conditional logic. *Theoria*, 36:23–42, 1970.
38. Johan F. van Benthem. *Modal correspondence theory*. PhD thesis, Mathematical Institute, University of Amsterdam, Amsterdam, NL, 1976.
39. Cornelis J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29:481–485, 1986. Also reprinted in [36], pp. 268–272.
40. Peter Willett, editor. *Document retrieval systems*. Taylor Graham, London, UK, 1988.
41. S.K. Michael Wong and Yiyu Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, 1995.